ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
UNIVERSITY OF PIRAEUS

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης
Μεταπτυχιακό Πρόγραμμα Σπουδών στην Εφαρμοσμένη Στατιστική

Κανόνες Διακοπής Κλινικών Δοκιμών Φάσης ΙΙΙ

Ανασκόπηση και Σύγκριση

Χαριτίνη Ανδριακοπούλου

Επιβλέπων καθηγητής: Σωτήριος Μπερσίμης

Διπλωματική εργασία που υποβλήθηκε στο τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική.

Πειραιάς, 2021

University of Piraeus

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από το ΓΣΕΣ του τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμόν ..... συνεδρίαση του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική.

Τα μέλη της επιτροπής ήταν:

- Αναπληρωτής καθηγητής Μπερσίμης Σωτήριος (Επιβλέπων)
- Αναπληρωτής καθηγητής Πολίτης Κωνσταντίνος
- Αναπληρωτής καθηγητής Τζαβελάς Γεώργιος

Η έγκριση της Διπλωματικής Εργασίας από το τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

ity of Piraeus University

Περίληψη

Η παρούσα διπλωματική εργασία παρουσιάζει μεθόδους για τον πρόωρο τερματισμό κλινικών δοκιμών φάσης ΙΙΙ και άλλων δοκιμών μεγάλου μεγέθους. Έμφαση δίνεται σε σχεδιασμούς στους οποίους οι ασθενείς εισέρχονται στη μελέτη σε διαφορετικούς χρόνους και αναλύονται κατά ομάδες σε μία σειρά επαναλαμβανόμενων στατιστικών συγκρίσεων που επιτρέπουν τη συνέχεια διενέργειας της μελέτης ή επιβάλλουν τον πρόωρο τερματισμό της.

Αρχικά θα εξετασθεί ο απλός σχεδιασμός ενός σταδίου στον οποίο υπάρχουν μόνο δύο ομάδες ασθενών, και η απόκριση μετριέται είτε με συνεχείς μεταβλητές ή με διάστημα χρόνου πρωτού συμβεί το αναμενόμενο γεγονός. Έμφαση θα δοθεί στο μαθηματικό υπόβαθρο του κάθε σχεδιασμού το οποίο παρουσιάζεται αναλυτικά για τους σχεδιασμούς που έχουν επιλεχθεί από τη βιβλιογραφία.

Στη συνέχεια θα εξετασθούν σχεδιασμοί δύο σταδίων οι οποίοι μπορεί να περιλαμβάνουν δύο ή και περισσότερες ομάδες ασθενών, καθώς και ενοποιημένοι κλινικοί σχεδιασμοί φάσης II/III. Οι τελευταίοι επιλέγονται κυρίως για λόγους εξοικονόμησης χρόνου (λ.χ. όταν υπάρχει άμεση ανάγκη για χορήγηση άδειας κυκλοφορίας στην αγορά), λόγοι κόστους καθώς και μείωση του αριθμού των ασθενών. Τέλος θα εξετασθούν σχεδιασμοί στους οποίους η απόκριση μετριέται με δύο μεταβλητές οι οποίες αξιολογούνται ταυτόχρονα και έχουν την ίδια βαρύτητα.

Στο δεύτερο μέρος της μελέτης, θα επιχειρηθεί μία σύγκριση των σχεδιασμών σε σχέση με το συνολικό δείγμα που απαιτούν για να επιτευχθεί η αναμενόμενη στατιστική ισχύς, καθώς και την ικανότητά τους να αντιληφθούν μία διαφορά στην επίδραση των δύο θεραπειών, εαν αυτή υπάρχει, εμπλέκοντας όσο το δυνατόν λιγότερους ασθενείς.

Οι κανόνες τερματισμού είναι σημαντικό μέρος του σχεδιασμού, πρωτίστως για λόγους ηθικής και σεβασμού προς τον ασθενή και την παγκόσμια κοινότητα. Το κλείσιμο της μελέτης επιβάλλεται σε περίπτωση που η νέα θεραπεία δεν είναι αποτελεσματική (ή δεν είναι το ίδιο αποτελεσματική όσο η ενδεδειγμένη) ώστε να εκτεθούν όσο γίνεται λιγότεροι ασθενείς και για λιγότερο χρονικό διάστημα, ή όταν η νέα θεραπεία διαπιστωθεί ότι είναι πιο αποστελεσματική από την ενδεδειγμένη ώστε άμεσα να γίνει διαθέσιμο το φάρμακο στην αγορά.

University of Piraeus

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
UNIVERSITY OF PIRAEUS

School of Finance and Statistics

Department of Statistics and Insurance Science
Master of Science in Applied Statistics

Stopping Rules in Phase III Clinical Trials

Review and Comparison

by

Charitini Andriakopoulou

Supervisor: Prof. Sotirios Bersimis

Submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics.

Piraeus, 2021

University of Piraeus

Author's statement:

I declare that this thesis was composed by myself under the guidance of my supervisor, that the work herein is my own except where indicated by referencing. When alluding to wording that' s not my own, a citation identifying the author and the full name of the source has been added. Where a source' s words were presented verbatim, italic style text has been used along with referencing of the source.

Of note, all the trial designs and the treatment schemes presented here are not real.

ity of Piraeus University

Quoted dialogue from *Multiple Testing in Clinical Trials, Statistics in Medicine, vol. 10, pp. 871-890 (1991)*

Mr. Louis: *"In the stagewise procedure, where you' re ordering p-values, for example, and then rejecting a hypothesis if the lowest one is lower than $\alpha/3$, and the next if $\alpha/2$, and so on. I understand that this strategy will protect the multiple level. But how do we interpret the rejections? We can say that we have rejected the hypothesis associated with the third smallest observed p-value, but that's not necessary the same thing as saying we reject a specific hypothesis. Could you discuss a bit how one takes the next step in terms of saying that this rejection applies to specific hypothesis? I may have made this more confusing than necessary."*

Mr. Bauer: *"I have no good answer for that other than the suggestion to report the individual p-values, which could be used in a descriptive way. One applies such multiple test procedures in order to satisfy the confirmatory aspect in terms of error probabilities, which is certainly a concept not unanimously agreed on by biometricians."*

ity of Piraeus University

Abstract

This review presents various methods to early terminate phase III and other large-scale clinical trials. Focus is given on the group sequential design which involves sequential patient enrollment and analyses of accumulating data at pre-defined inspection times. Efficacy evaluation is performed in the context of a repeated significance testing to allow for early stopping.

In the first part, several approaches for the one-stage, two-arm design will be discussed. Treatment efficacy is being evaluated using either a continuous or a time-to-event endpoint. Emphasis is given in the mathematical background of each approach.

Subsequently, various approaches for the multi-stage design will be discussed, as well as the seamless phase II/III design. The latter is widely used for time saving reasons (e.g. due to an urgent need for drug approval), for cost-reducing reasons and for minimising the sample size. Designs having two co-primary endpoints will also be discussed.

In the second part, a comparison among various designs is attempted with regard to the total sample size needed to reach the expected statistical power, and their ability to detect a treatment difference, if any, without allowing a large number of participants to be engaged.

Stopping rules constitute an integral part of the statistical design, primarily for ethical reasons. Early termination must take place in case of the new treatment proven to be inefficient (or not as efficient as the existing one) to prevent more patients from being exposed to ineffective drugs. Early stopping must also take place in case of the new treatment proven to be superior to the existing one to rapidly make it available to the public. Other reasons for stopping a trial, such as high toxicity rate, low accrual rate, arising breakthrough treatments, are out of scope of this review.

ity of Piraeus University

Contents

University of Piraeus

Glossary

clinical benefit A favourable effect on a meaningful aspect with regard to patient's clinical condition. Clinical benefit may refer to symptom relief, functionality or survival rate. A treatment may induce a pathological response (e.g. tumor size reduction) without having a clinical benefit (e.g. increased survival).

clinical relevance (also known as clinical significance) Indicates whether the results of a study are meaningful with regard to the disease under research. Should not be confused with statistical significance.

comparative study A clinical trial with two or more arms with the intention to compare between them. Clinical trials can have several arms with no intention to compare (non-comparative studies).

confirmatory trial An adequately controlled trial with a well-defined research question and a clinically relevant endpoint. Confirmatory trials intend to provide firm evidence of efficacy and may lead to drug licensing.

endpoint A variable that can be measured objectively to determine whether the intervention being studied is beneficial.

event An event occurs when the endpoint under study is reached. For instance, if the endpoint is overall survival, event is death.

interim analysis Pre-planned, statistical analysis of accumulated data with regard to treatment efficacy (safety data are also considered) to early terminate a trial in favour or against the experimental treatment. Other vocabulary used here with a similar meaning: *inspection time*, *look* and *group*.

information time The ratio of the number of evaluated patients to the total number of patients planned to be evaluated. Another definition: the number of events at a specific time point t over the total number of events needed to reach the expected power.

ity of Piraeus University

Brief history

Drug development dates back to the early days of human civilization - from the Chinese traditional medicine originated in 3500BC to Indian, Egyptian, Roman and Greek medicine which is traced back to 400BC. Hippocrates the "Father of Medicine", founded a medical school and established the first ethics for physicians. Early substances derived from plants, animals and minerals, were developed through observation and error experimentation on humans and animals. [1]

The Renaissance period marked the transition from Middle ages to modern world and besides others, laid the foundation for medicinal science as we know it today. In the late 1700s, Edward Jenner worked with smallpox inoculations leading the way for vaccine development. John Hunter (1768) discovered that scurvy is caused by lack of vitamin C. Louis Pasteur (1864) devised a vaccine against rabies and saved thousands of lives. More recently, Alexander Fleming discovered the potency of Penicillium mold against staphylococcus. [1]

Synthetic drugs, the ones that we use today, were developed in the early 1900s, at the time when the pharmaceutical industry was founded. [1] *By early 1940s, new arrivals in drug industry, such as sulfanilamide, amphetamines and insulin made the need of organised, well-performed clinical trials more urgent.* [2] *The first double blind clinical trial took place in 1943 investigating patulin for the treatment of common cold. Although, patulin turned out to be ineffective, the trial pioneered as the first controlled human trial.* [3]

Modern history of human experiments is long and sombre as much unethical research has been carried out in the name of science. Since the late 19th century, several notorious experiments had taken place including exposing people to pathogens, toxic and radioactive chemicals, radiation and other deleterious agents. Much of the research was conducted on children, prisoners and mentally disabled people. [4] *Between 1946 and 1948, a large-scale experiment of approximately 1,500 highly vulnerable people was carried out in Guatemala. People were infected with gonorrhea and syphilis under the guise of STD prevention. The study was funded by the US. Public Health Service and was never published.* [5] *During World War II, there is strong evidence*

University of Piraeus

that Nazis conducted hundreds of experiments on war prisoners which mostly ended in pain, physical and phycological trauma, amputation and often death. Humans were forced to participate or they were deceived. A total of 23 officials, doctors and administrators were prosecuted for crimes against humanity in what today is known as the Doctors' Trial in Nuremberg. [6]

Guidelines

In 1949, following the Nuremberg's trials, a list specifying 10 conditions under which medical experiments can be allowed was published as the Nuremberg's Code. Among others, the Code emphasises the importance of voluntary consent, risk to human subjects, animal experimentation preceding human trials, right to withdraw, and early termination. [7] *Even though it does not have any legal status, the Nuremberg's Code formed the foundation on which later regulations were based. In 1964, in their annual meeting, the World Medical Association adopted the Declaration of Helsinki.* [8] *Ever since, the Declaration has been revised many times with the most recent major revision taking place in 2013. The Declaration of Helsinki is the basis of Good Clinical Practice (GCP) and other ethics guidelines issued worldwide. It highlights the need for written consent, restricts the use of placebo, and introduces the concept of prior review and approval of protocol by independent institutional review boards.*

More recently, different laws, guidelines and requirements among countries highlighted the importance of harmonisation. The inception of the International Council for Harmonisation of technical requirements for pharmaceuticals for human use (ICH) was born in 1990 when industry associations and regulatory agencies of Europe, Japan and the US met in Brussels. These harmonisation guidelines would eventually become the basis for developing, testing and authorising medicinal products. Today the guideline ICH E6 (commonly known as the ICH GCP guideline) is considered the most important guideline that covers clinical trials and constitutes a joint standard for mutual acceptance of clinical data by regulatory authorities. [6][9]

Among others, ethical and regulatory requirements suggest that sample size shall be large enough to assure robustness, but not larger than required as this would undermine beneficence of participants. Controls and suitable safeguards shall be in place to ensure that trials will be stopped after statistical significance is reached, or futility will be shown early enough without allowing further enrolment and treatment continuation.

Introduction

1.1 Modern clinical trials

Clinical trials are used to evaluate the effects of new treatments in the human health in terms of efficacy, clinical benefit, quality of life and side effects. They are carefully designed, reviewed and completed to answer 3 main questions: *Does the treatment work? Does it work better than the existing one? Does its overall benefit overtake its harm?*

According to the ICH, a clinical trial is *any investigation on human subjects intended to discover or verify the clinical, pharmacological or other pharmacodynamic effects of an investigational product, to identify any adverse reactions related to the investigational product, to study absorption, distribution, metabolism, and excretion with the object of ascertaining its safety and/or efficacy.* [9] People consent to take part in the clinical research to test medical interventions including drugs, medical devices, cells and other biological products, surgical procedures, radiological procedures, preventive care and other.

Each trial is strictly conducted according to a comprehensive plan: the protocol. The protocol outlines the accrual, the treatment scheme, the schedule of tests and procedures, the follow-up period and the length of the study. It also describes the variables that will be measured and the type of information that will be collected.

Pre-market clinical research consists of three phases examining safety and efficacy of the new treatment. In the first phase, toxicity profile and dose limiting toxicity are explored; efficacy may be recorded as well, but it does not constitute an endpoint (no conclusions on efficacy shall be drawn). Phase I trials are almost always non-randomised and typically terminate after several months. Sample size typically ranges from 20 to 100 participants (healthy or having the disease/condition), and approximately 7/10 drugs move to the next phase. [10]

In the second phase of clinical research, proving efficacy is the main goal whilst

toxicity rate is constantly recorded. Sample consists of up to several hundred people having the disease/condition, trials typically last several months to two years and approximately 1/3 drugs move to the next phase. Caution should be taken when interpreting results as sample size can be relatively small and effective drugs may be rejected or vice versa. [10]

In the last pre-market clinical phase, research focuses on showing equivalence, superiority or non-inferiority of the investigational product as compared to standard treatment, best supportive care, or placebo (if this is the case). Phase III trials are always randomised and comparative, consist of 300 to 3,000 participants having the disease/condition, and typically last one to four years. Approximately, 25−30% of drugs entering phase III get market approval. [10]

Clinical research also takes place in the post-market surveillance with sample size being as large as the entire population receiving the treatment (phase IV follow-up studies). Post-market surveillance may lead to licence recall and drug removal from the market if unexpected toxicity or poor real-world clinical outcome is recorded.

Traditional clinical trials have two main drawbacks - there are time consuming and often need a significant number of participants to obtain robustness. In industry's world, fast standardised trials denote rapid drug registry and low cost. In patients' world, often it is a matter of life and death. In science's world, designing flexible and time-saving trials is always a challenge. Recently, the adaptive seamless design has gained ground. This approach combines two or more phases in a single trial and allows for real-time design modifications. Seamless trials are flexible, cost-saving and often lead to speedy approval procedures.

1.2 Key issues in the design phase

Clinical research is a highly interdisciplinary field with focus on the human. Before any intervention, the primary research question shall be addressed; primary endpoint, eligibility criteria, treatment scheme, randomisation and statistical approach are all selected in such a way so as to account for the primary objective of the study. In confirmatory trials, a suitable, clinically relevant endpoint is evaluated to answer a specific pre-defined question, whilst other questions may be explored by secondary endpoints. In some cases, more than one primary endpoints (often correlated) are used to answer the research question, with statistical design addressing the issue of multiplicity and endpoint correlation.

The parameter of interest broadly depends on the nature of endpoint used to assess

efficacy. For instance, if the measured response is a continuous variable (e.g. sugar levels), a suitable parameter to construct the test statistics would be the means difference, whilst in a clinical trial involving failure times as seen in cancer research, a suitable parameter would be the log-hazards ratio or a Kaplan-Meier estimate of the median survival time.

1.3 Repeated significance testing

A fixed sample size design demands that *the results of the experiment will be judged once and for all when the (pre-defined) sample size is reached.* [11] However, ethics requires minimum sample size and early termination, especially when the new treatment is inferior to the standard of care. For this reason, the biometrician is asked to analyse accumulating results at pre-planned time points.

It is widely known that performing sequential tests inflates type I error−10 repeated tests increase α from 5% to 19%, and if one continues testing, one can be certain of the existence of a treatment difference whether one is present or not. [11][12] To overcome this burden, the global level α has to be preserved. The concept of repeated significance testing was born in the late 1960s by Armitage, McPherson & Rowe. [13][14] It involves choosing appropriate decision regions such as the actual significance level does not exceed the global α. This is feasible if a fixed schedule of inspections has been decided before the initiation of the trial.

Repeated significance testing constitutes an early termination method; it is often used in the context of a group sequential design which is widely used in oncology and other disease areas. In the group sequential approach, response is tested at a series of interim analyses on accumulating data maintaining an overall type I error. This approach is commonly used in oncology and other disease areas where patients are gradually involved and time before an event occurs may be of weeks or up to several months. Group sequential and adaptive group sequential designs will be extensively discussed throughout this review. Table 1.1, an example by Pocock (1982) [12], presents a group-sequential design with two treatment arms for various numbers (1-4) of maximum interim analyses (groups). Expected and average sample size have been determined through numerical integration. Note that the sequential design slightly increases the maximum number of needed patients while reduces the average number when treatment difference actually exists.

Joint distribution of the test statistic

In the context of a group sequential design, interim analyses take place at regular time points with the test statistic of each arm being compared to a stopping boundary. It is important to realise that the joint distribution of test statistic computed in the context of repeated significance testing only depends on the number of observations (or events) at the looks to date (even though it seems to depend on the total sample size or the needed events to reach the expected power). [15] This is because the joint distribution depends only on the ratios of the information times ($t_i/t_j = n_i/n_j$) with n_i denoting the group size and therefore, the biometrician does not really need to know in advance the information times $(t_1, t_2, ..., t_j)$ of the first j looks to determine the joint distribution of the test statistic up to point j. [16]

J	α^*	$2n$	$2nJ$	Average $2nJ$ under H_1
1	0.05	52.0	52.0	52.0
2	0.029	28.4	56.8	37.2
3	0.022	19.7	59.2	33.7
4	0.018	15.2	60.8	32.2

Table 1.1: Group sequential designs for a normal response with known variance σ^2, global $\alpha = 0.05$, power $1 - \beta = 0.95$, alternative hypothesis $H_1 : \mu_A - \mu_B = \delta$ and several choices of maximum number J of groups. Table presents the nominal level α^*, the required number of patients per group for both arms $(2n)$, the maximum number of patients $(2nJ)$, and the average number of patients until trial termination under H_1. Sample sizes have to be multiplied by σ^2/δ^2. [12]

1.4 Multiple testing

Multiple testing is often performed when more than one treatment is compared with the control, when the trial involves more than one primary endpoint, or in an adaptive scheme combing several stages. The main concern of such a design is how to control the global level α, or otherwise referred to as the *family error*. In his work *"Multiple Testing in Clinical Trials"*, Bauer states that the reason our interest is restricted in type I error is because one would give up an existing standard of care in favour of a new promising treatment only when strong evidence demonstrates that one should do so. In real world practice, we often have far less knowledge on the arising treatment compared to the existing one. This fact alone creates an inherent imbalance between the two treatments rendering risks of erroneous decision hardly foreseeable. [17]

But even if one accepts the type I error philosophy, is it the expected number of erroneous rejections or the probability of at least one erroneous rejection that should be controlled? [18] Bauer believes that for clinical trials involving a small number

of null hypotheses to be decided on (which is often the case), it seems reasonable to control the probability of at least one erroneous rejection, also controlling the probability of not performing any erroneous rejection at all. [17] In the conventional approach, this probability is controlled under the global null hypothesis which states that all individual hypotheses are simultaneously true:

$$H_o = \bigcap_{i=1}^{k} H_{oi}$$

A multiple test procedure controls the multiple level α (the family error in the strong sense) if the probability of erroneously rejecting at least one true individual null hypothesis is controlled by α, irrespective of which and how many of the individual null hypotheses are in fact true. [17]

Even though there is much more to be said for multiple testing, further exploration is out of scope of this review. The interested reader may refer to work by Paulson (1962, 1964) [19][20], Dunnett (1965) [21], Bechhofer et al. (1968) [22], Armitage et al. (1969) [13], Spjøtvoll (1972) [23], Holm (1979) [24], DeMets & Ware (1980) [25], Hsu & Edwards (1983) [26], Simes (1986) [27], Kim et al. (1987) [28], Hochberg et al. (1987, 1988) [29][30] and Bauer (1987–1989) [31–35]. Some of this work will be discussed here focusing on stopping rules of these designs.

1.5 Stopping by design

Stopping rules are introduced before initiation of the trial, in the design phase. They prevent researchers from involving a large number of participants, and also they save time and cost. Ineffective drugs will be rejected in the initial stages of the trial, without allowing a large number of patients to be exposed to them, whilst effective treatments will rapidly manage to get though the licence process and become available to the public. In phase I designs with primary endpoint drug toxicity, stopping rules often aim at early termination when serious adverse events are recorded. In phase II & III designs, early stopping takes place if drug effectiveness or ineffectiveness is proven, although unexpected toxicity is always a reason to stop.

Repeated significance testing is often performed to early terminate a trial with sequential patient enrolment. The total sample size is not fixed; the study continues until the test statistic ξ_j exceeds its specified boundary. Common choices for ξ_j include the maximum likelihood estimate, a normalised Z statistic based on H_o, a p-value, Bayesian posterior probabilities, Bayesian predictive probabilities, conditional power etc. [36] The maximum number of interim analyses and the group sample size

are often determined in advance. Calculations for the latter are based on parameter estimates (usually from historical data) and expected statistical power. At each look, the outcome space for ξ_j is partitioned into two regions: the stopping and the continuation region. One could have more stringent significance level at the first looks of the trial (to prevent early termination without having a substantial treatment difference) and less stringent later on. [37] The most important aspect of a group sequential design is its ability to detect a treatment difference soon enough to allow for early stopping. This aspect is reflected through the last column of Table 1.1: the average number of patients under the alternative hypothesis.

During the course of the trial, it is very common to deal with unexpected issues such as low accrual rate, low event rate, extreme censoring or treatment deviations. Usually these issues may lead to protocol amendments or even trial discontinuation (e.g. due to unexpected serious adverse events or arising breakthrough treatments). Stopping the trial due to unacceptable toxicity is planned in advance, however it is highly dependent on the nature of adverse events that will arise. Stopping for toxicity is out of scope of this review.

1.6 Aim and scope

The aim of the present review is to present and discuss broadly accepted designs to early terminate large-scale randomised clinical trial. In the first part, the mathematical background of some widely used designs for the group sequential approach and various adaptive designs for two-arm and multi-arm trials is presented. The reader may refer to relevant sources to further explore these methods.

In real world, choosing a suitable design broadly depends on factors such as the disease area, the question to be answered, the primary endpoint, the available number of patients, the resources, and even how urgent the need for market approval is. Some designs perform better than others in terms of sample size, time and cost. In the second part, we attempt to compare various designs in terms of the total sample size needed to reach the expected statistical power, and their ability to detect a treatment difference, if one exists, without allowing a large number of participants to be engaged.

Group sequential design

2.1 Introduction

The strict sequential design for a two-arm trial requires that patients are enrolled in pairs, and that a pair can only enter the trial after the results of the prior pair have been analysed. This costly and time-consuming design is not practical, especially when a large number of patients needs to be engaged (e.g. in phase II&III trials). In 1947, Wald [38] suggested a grouped analysis for the sequential design where testing is performed once a certain number of patients have been involved. Group sequential test procedures are actually quite older since they were used for quality control already from the late 1930s [39] (for a historical review see [40]). Today, the group sequential design is the gold standard when patients enter the trial at different time points and endpoint assessment is not done instantly.

The group sequential design is widely used in clinical research, especially in phase II&III trials. Accumulating data is analysed at pre-defined time points (usually in terms of number of events, or information time) and early termination takes place when the ultimate decision is known with high confidence. More precisely, stopping may occur in case of treatment superiority, non-inferiority or equivalence (depending on the trial's objective; all of them referred here as *stopping for efficacy*) or in case of foreseeable failure to reject the null hypothesis (*stopping for futility*). One of the benefits of this design is the possibility that it offers to conduct repeated significance tests having a fixed number of patients (or events) per group; the latter has to be determined in advance, independently of the acquired data. [41]

In certain cases, fixed group sample size may turn out to be a burden, for example due to unforeseeably slow accrual or event rate. Fortunately, in the context of an adaptive design, the expected number of observations per group may be changed as the trial goes on. The concept of the adaptive design was due to Bauer (1989) [35] and Bauer & Köhne (1994) [42]. Today, adaptive designs allow for real-time modifications such as sample size recalculation, adding or dropping treatment arms (e.g. drop the loser design), modifying the treatment scheme, interfering with the

University of Piraeus

randomisation method (e.g. adding or removing stratification factors), a combination of the aforementioned changes or other adaptations, all in line with the protocol specifications.

In the applied field, biometricians often prefer to perceive the problem of sequential testing identical to that of multiple testing. Having multiple primary endpoints being tested simultaneously is viewed in a similar manner as having one (or multiple) endpoint being assessed at multiple consecutive times (as the same null hypothesis would be tested repeatedly). Hence, a sequential test of global level α can be treated as a test of a single hypothesis using a multivariate test statistic. The main difference between a multiple testing scheme and a sequential design lies in the fact that in the latter, a stopping rule must be used to early terminate the trial if needed. [17]

Early work on the group sequential design has been carried out by Popock (1977) [43], O'Brien & Fleming (1979) [37], Lan & Demets (1983) [44], Kim & DeMets (1987) [45] and others. Much of this work referred to a previous method proposed by Armitage et al. (1969) [13], subsequently known as recursive numerical integration (or Simpson's rule). Recursive numerical integration is a quite sophisticated method used to estimate the score statistics joint distribution allowing for early stopping. The method originally referred to a sequential design in which patients enrol in matched pairs and significance testing takes place instantly in a continuous monitoring scheme. However, as stated earlier, continuous monitoring is often non-feasible due to practical issues such as reporting delays and other time-consuming administrative procedures. [46] Since then, most trials with sequential enrolment used ill-defined rules for early stopping. [43]

In 1977, Pocock [43] modified recursive numerical integration to allow for repeated significance testing of accumulating data. He discussed two-sided tests with variables having a normal response and known variance whilst assuming equal number of observations between looks. Significance level, as well as stopping boundary, at each interim remains the same. Since then, several approaches have been suggested with fewer restrictions. O' Brien & Fleming (1979) [37] used a consecutive test statistic with variance proportional to sample size so that stopping in favour of the alternative is less likely to occur in the beginning of the trial, where a small accumulated sample size would most probably be insufficient to a prove treatment difference. Both methods ([43][37]) assume equally spaced looks (in terms of information time). A couple of years later, Pocock investigated a sequential design with varying nominal significance levels. [12] Wang & Tsiatis (1987) [47] proposed a family of critical values with Pocock's and O' Brien & Fleming's being sub-cases.

A different approach by Slud & Wei (1982) [46] suggested discrete sequential bound-

aries based on modified-Wilcoxon scores (Gehan scores). At the jth of J looks (J specified in advance), a part π_j of the global level α is spent so that $\sum_{j=1}^{J} \pi_j = \alpha$. Slud & Wei's method can be applied to other statistics as well (e.g. Tsiatis (1982) [48]). [16] It can be applied even if the inspection times are not specified in advance (though the maximum number of looks has to be pre-specified). [49]

Even though defining in advance the maximum number of inspection times is a tempting approach, and actually constitutes the gold standard in many disease areas, in some cases the investigator might simply not have a good estimate for that. In 1983, Lan & DeMets [44] proposed the use of an α-spending function without having to specify in advance the maximum number of interim looks, just the total sample size. They used an increasing function $\alpha(t)$ with t denoting the information time, $\alpha(0) = 1$ and $\alpha(1) = \alpha$ so that at each interim analysis j, only a part $(\alpha(t_j) - \alpha(t_{j-1}))$ of global α is spent. In 1986, Bauer [50] used Lan & DeMets's approach and suggested a design that neither the maximum number of interim analyses nor the total sample size have to be specified in advance. His method only requires the joint distribution of the test statistics at the last two inspection times. Also, the Fleming, Harrington & O' Brien's approach (F-H-O method) [49] allows for increasing the number of interim analyses if needed (e.g. due to slow accrual or event rate).

Stochastic curtailment is another method that does not require to specify in advance the number of interim looks. In a different line of thought, this approach is based on data accumulated up to a point t and the null hypothesis for the remaining time $T - t$. Using this method, the biometrician projects ahead and computes the probability of an H_o rejection at the end of the study. [51] Bayesian methods also do not require to specify in advance the number of interim looks. A prior distribution for the parameter of interest θ is specified and continuously updated as more data comes in, whilst decision on trial termination is based on the posterior distribution.

In 1992, Proschan, Follmann & Waclawiw [16] in a very interesting publication, investigated the degree to which assumption violations can inflate type I error in several sequential monitoring schemes. Among others, they examined violation of the assumption of equal spacing (in terms of information time) between interim analyses, and to what extent type I error is affected in the designs where future looks are planned based on data trends.

In this chapter, focus is given on the mathematical background of selective group sequential approaches for comparing a single experimental treatment with a control. Tables with expected number of patients and statistical power are provided along with methods as reported by the authors.

ity of Piraeus University

2.2 Recursive numerical integration (Simpson's rule)

Consider the problem of comparing an experimental treatment with a control whilst having a continuous endpoint. The score statistics S_j (with j denoting the interim analysis) could be expressed as follows:

$$S_j = \hat{\theta}_j \mathscr{F}_j (j = 1, ..., J) \tag{2.1}$$

with θ denoting the difference in efficacy between the two treatments, and $\mathscr{F}_j$ the Fisher information table. Under H_o, S_j follows the multivariate normal distribution with:

$$S_j \sim N(\theta \mathscr{F}_j, \mathscr{F}_j) \tag{2.2}$$

so that the quantity $S_j - S_{j-1}$ is independent of $S_1, ..., S_{j-1}$.

At interim analysis j:

- if $S_j \geq u_j$, the trial is stopped and H_o is rejected
- if $S_j \leq l_j$, the trial is stopped and H_o is not rejected
- if $l_j < S_j < u_j$, the trial continues to interim analysis $j + 1$

The stopping boundaries $(l_1, u_1), ..., (l_J, u_J)$ are chosen so as to control the overall type I error. For the exact test:

$$P(stop\ and\ reject\ H_o \mid \theta = 0) = \alpha \tag{2.3}$$

Many choices of $(l_1, u_1), ..., (l_J, u_J)$ will satisfy (2.3), and therefore critical values will also have to satisfy another condition: an α-spending function. For the one-sided test, boundaries for efficacy/futility stopping can be computed using two conditions [52]:

$$P(stop\ and\ reject\ H_o\ at\ or\ before\ interim\ analysis\ j \mid \theta = 0) = \alpha_U^*(t_j)$$

$$P(stop\ and\ do\ not\ reject\ H_o\ at\ or\ before\ interim\ analysis\ j \mid \theta = 0) = \alpha_L^*(t_j) \tag{2.4}$$

where

$$t_j = \mathscr{F}_j / \mathscr{F}_J, \quad \alpha_U^*(0) = 0, \quad \alpha_U^*(1) = \alpha, \quad \alpha_L^*(0) = 0 \quad and \quad \alpha_L^*(1) = 1 - \alpha$$

The joint distribution of $(S_1, ..., S_J)$ under H_o can be approached using recursive numerical integration as follows. According to (2.2), under H_o:

$$S_1 \sim N(0, \mathscr{F}_1)$$

thus,

$$f_1(s) = \phi(s/\sqrt{\mathscr{F}_1})\,\frac{1}{\sqrt{\mathscr{F}_1}}$$

where ϕ denotes the standard normal density function. To satisfy (2.4), critical values (l_1, u_1) are set to the lower $\alpha_L^*(t_1)$ and upper $\alpha_U^*(t_1)$ points of this distribution. S_2 follows a distribution having a sub-density of S_1 distribution; this sub-density is equal to the area within the continuation region of S_1 with $S_2 \in (l_1, u_1)$. Accordingly, the density of S_j distribution equals the continuation region of S_{j-1} with $S_j \in (l_{j-1}, u_{j-1})$ (see figure 2.1; left panel). $S_j - S_{j-1}$ is normally distributed and independent of S_{j-1} so that:

$$f_j(s) = \int_{l_{j-1}}^{u_{j-1}} f_{j-1}(s_{j-1})\frac{1}{\sqrt{\mathscr{F}_j - \mathscr{F}_{j-1}}}\,\phi\left(\frac{s - s_{j-1}}{\sqrt{\mathscr{F}_j - \mathscr{F}_{j-1}}}\right) ds_{j-1}, \quad j = 2, ..., J \quad (2.5)$$

translates into the probability of continuing to the j interim analysis. Critical values (l_j, u_j) are set to the lower $\alpha_L^*(t_j) - \alpha_L^*(t_{j-1})$ and upper $\alpha_U^*(t_j) - \alpha_U^*(t_{j-1})$ points of $S_j - S_{j-1}$ density, respectively.

Recursive numerical integration is an alternative technique to a rather ill-defined approach of repeated significance tests to accumulating data which inflates type I error and increases the possibility of erroneously rejecting the null hypothesis. [43] The method constitutes a milestone in the group sequential statistical design, and over the years it has been discussed and adjusted by multiple authors.

2.3 Alpha-spending functions

In bibliography, various forms for the spending functions $\alpha_L^*(t)$ and $\alpha_U^*(t)$ exist, with one of the most popular being that of Lan & Demets's [44]. In 1983, Lan & DeMets suggested a method to compute a flexible discrete boundary $(u_1, ..., u_J)$ using a predefined increasing α-spending function $\alpha^*(t)$ without having to decide in advance the number of maximum looks J. Three possible choices of $\alpha^*(t)$ were introduced:

ity of Piraeus University

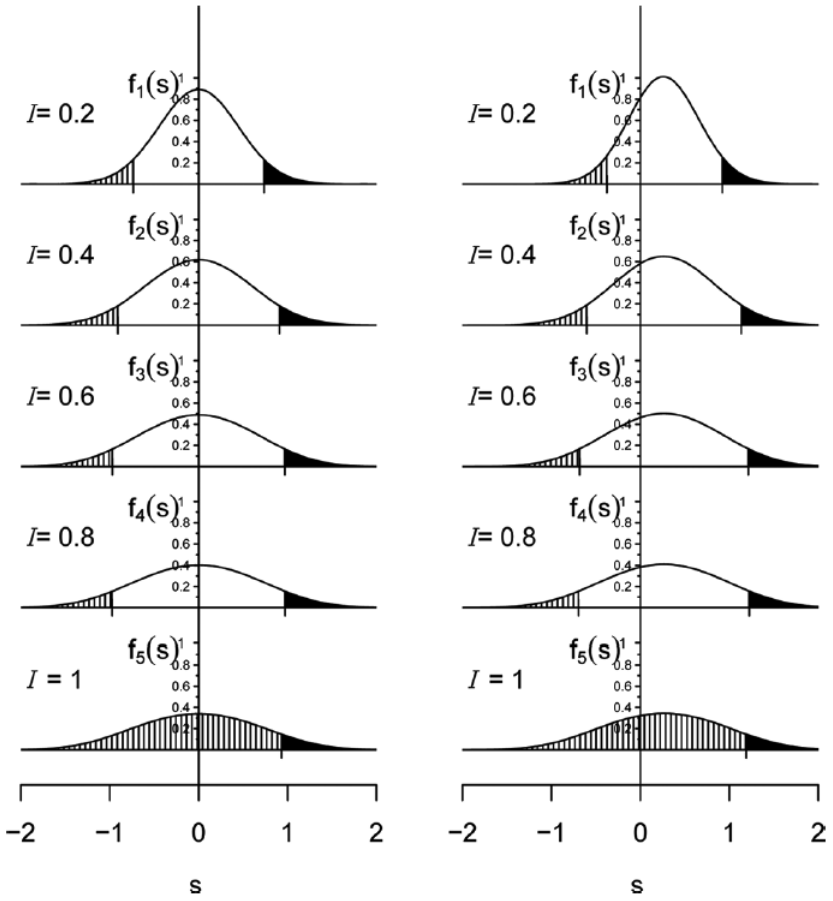


Figure 2.1: Plots of sub-densities $f_1(s), ..., f_5(s)$ to compare a single experimental treatment with a control (left panel) and to select the best of 3 experimental treatments at the first interim analysis (right panel) for a trial with 5 analyses at $\mathscr{T} = 0.2, ..., 1$. Areas corresponding to stopping with H_o rejection are heavily shaded; areas corresponding to stopping for futility are lightly shaded; areas corresponding to continuation are unshaded. The total area under each sub-density is equal to the area within the continuation region under the sub-density for the previous look. [53]

	$\alpha = 0.025$					$\alpha = 0.05$				
	c_1	c_2	c_3	c_4	c_5	c_1	c_2	c_3	c_4	c_5
Pocock	2.41	2.41	2.41	2.41	2.41	2.12	2.12	2.12	2.12	2.12
O' Brien & Fleming	4.56	3.23	2.63	2.28	2.04	3.92	2.77	2.26	1.96	1.75
Lan & DeMets - $\alpha_1^*(t)$	4.90	3.35	2.68	2.29	2.03	4.23	2.89	2.30	1.96	1.74
Lan & DeMets - $\alpha_2^*(t)$	2.44	2.43	2.41	2.40	2.39	2.18	2.14	2.11	2.09	2.07
Lan & DeMets - $\alpha_3^*(t)$	2.58	2.49	2.41	2.34	2.28	2.33	2.22	2.12	2.03	1.96

Table 2.1: One-sided boundary for the standard Brownian motion at 5 interim analyses using Pocock's method, O' Brien & Fleming's method and three different approaches by Lan & DeMets. [44]

$$\left.\begin{aligned}\alpha_1(t) &= 2 - 2\,\Phi\left(\frac{z_{\alpha/2}}{\sqrt{t}}\right)\\ \alpha_2(t) &= \alpha\, log\,(1 + (e-1)\,t)\\ \alpha_3(t) &= \alpha t\end{aligned}\right\} \tag{2.6}$$

with $0 < t \leq 1$, $\alpha_1(0) = \alpha_2(0) = \alpha_3(0) = 0$ and $\alpha_1(1) = \alpha_2(1) = \alpha_3(1) = \alpha$.

H_o is rejected at information time t_j if:

$$Z(t_j) > C(t_1, ..., t_j) \tag{2.7}$$

with $C(t_1, ..., t_j)$ chosen so that:

$$P\left(Z(t_1) \leq C(t_1), ..., Z(t_{j-1}) \leq C(t_1, ..., t_{j-1}), Z(t_j) > C(t_1, ..., t_j)\right) = \alpha(t_j) - \alpha(t_{j-1}) \tag{2.8}$$

From (2.8) it is clear that the amount of α to be spent at the next look is integrally dependent on its information time. This is the reason why the number of looks do not need to be specified in advance. [16]

Table 2.1 presents stopping boundaries for 5 interim analyses using Pocock's [43] and O'Brien & Fleming's [37] methods, and the aforementioned α-spending functions.

Another work by Hwang, Shih & De Cani (1990) generalised Lan & DeMets's $\alpha^*(t)$ to the following one-parameter truncated exponential distribution family of α-spending functions $\alpha^*(\gamma, t)$ [54]:

$$\alpha^*(\gamma, t) \;=\; \begin{cases} \alpha\dfrac{1 - e^{-\gamma t}}{1 - e^{-\gamma}}, & \gamma \neq 0 \\ \alpha t, & \gamma = 0 \end{cases} \qquad 0 \leq t \leq 1 \tag{2.9}$$

so that $\alpha^*(\gamma, 0) = 0$ and $\alpha^*(\gamma, 1) = \alpha$ for all γ. Time t denotes the information time.

2.4 Pocock's method

In 1977, Pocock [43] modified the recursive numerical integration method to allow for group sequential testing as the method was originally designed for simple sequential testing. Having two treatment arms A and B of n patients each, with normal responses $\bar{x}_A$ and $\bar{x}_B$ and a common known variance σ^2, after j interim analyses:

$$\bar{d}_j = \sum_{i=1}^{j} \frac{\bar{x}_{Ai} - \bar{x}_{Bi}}{j} \sim N\left(\mu_A - \mu_B,\, \frac{2\sigma^2}{j n_j}\right) \tag{2.10}$$

University of Piraeus

A two-sided significance test allows for the p-value of:

$$p_j = 2\left(1 - \Phi\left(\frac{\bar{d_j}\sqrt{jn_j}}{\sqrt{2\sigma}}\right)\right) \tag{2.11}$$

to be compared with some nominal significance level $\alpha^{'}$. If $p_j < \alpha^{'}$, the null hypothesis of $\mu_A = \mu_B$ is rejected and trial is stopped. Else, if $p_j \geq \alpha^{'}$, trial continues to the next interim analysis.

The whole idea behind Pocock's method is to allow for repeated testing while maintaining an overall type I error (α) and having pre-defined a maximum number of interim analyses. At each interim look, significance testing takes place at the same nominal level α^*, with α^* not depending on the sample size n. In practice, H_o is rejected if:

$$Z(t_j) > C_P^{(J)} \tag{2.12}$$

where $C_P^{(J)}$ is a constant chosen so that the overall probability of rejection under H_o (assuming equally spaced information time points $t_j = j/J$) is α.

Interestingly, Pocock's boundary can be approximated by a Hwang, Shih & De Cani's spending function (with $\gamma = 1$).

Table 2.2 provides the average number of patients under both H_o and H_1 for several values of maximum interim analyses J and for specific values of overall type I error (α) , expected treatment difference ($\mu_A - \mu_B$) and power ($1 - \beta$).

No. of maximum interim analyses	1	2	3	5
Required no. of patients per group and treatment arm	84.1	46.2	32.2	20.3
Maximum no. of patients per treatment arm	84.1	92.4	96.6	101.5
Average no. of patients per group and treatment arm under H_1	84.1	65.2	60.5	57.5

Table 2.2: Required, maximum and average (under the alternative hypothesis) number of patients for several numbers of maximum interim analyses in a group sequential design of two treatment arms with normally distributed response and known common variance σ^2 using Pocock's method ($\alpha = 0.05$; $\mu_A - \mu_B = 0.5\,\sigma$; $1 - \beta = 0.90$). [43]

2.5 O' Brien & Fleming's method

A couple of years later, O' Brien & Fleming [37] published an alternative approach for terminating the group sequential design in which α_j increases with j such that $\alpha_J \simeq \alpha$. This means that, unlike Pocock's method, O' Brien & Fleming's method allows for having a stringent stopping rule at the beginning of the trial and less stringent later on. This is due to the fact that stopping in favour of the alternative is less likely to be decided at the first looks where the small number of accumulated patients (or events) would most probably be insufficient to show any treatment difference. At each look, H_o is rejected if:

$$Z(t_j)\sqrt{\frac{j}{J}} > C_{O-F}^{(J)} \tag{2.13}$$

with $j \in \{1, 2, ..., J\}$ denoting the interim analysis. O' Brien & Fleming's boundary can be approximated by a Hwang, Shih & De Cani's spending function with $\gamma = -4$.

O' Brien & Fleming's bound is widely used in clinical trials utilising the group sequential design, especially in the analysis of censored survival data since it preserves the sensitivity to late occurring survival differences. [49] Table 2.3 gives $C_{O-F}^{(J)}$ for 5 different values of global α for 5 inspection times.

	Look (j)				
α	1	2	3	4	5
0.05	3.87	3.92	3.94	4.17	4.15
0.04	4.29	4.23	4.26	4.48	4.58
0.03	4.80	4.72	4.70	4.96	5.05
0.02	5.49	5.39	5.46	5.56	5.79
0.01	6.67	6.57	6.50	6.86	6.84

Table 2.3: Approximate values of stopping boundary $C_{O-F}^{(J)}$. [37]

2.6 Wang & Tsiatis's method

In 1987, Wang & Tsiatis [47] introduced a family of critical values with Pocock's ($\Delta = 0.5$) and O' Brien & Fleming's ($\Delta = 0$) boundaries being family members. Authors considered a trial with two treatment arms A and B, with a unknown normal response X and a known common variance σ^2. Rejection of $H_o : \mu_A = \mu_B$ can be managed if:

$$|S_j| \geq \alpha_j, \qquad j = 1, ..., J \tag{2.14}$$

with

$$S_j = \sqrt{n_j} \sum_{i=1}^{j} \frac{\bar{x}_{Ai} - \bar{x}_{Bi}}{\sqrt{2}\,\sigma} \tag{2.15}$$

with $\bar{x}$ denoting the mean response, and n_j the sample size at j interim analysis per treatment arm ($j = 1, ..., J$). Stopping boundaries α_j are chosen so that an overall α is maintained. The probability of failing to reject H_o while it is true is:

$$P(\,|S_1| < \alpha_1, ..., |S_J| < \alpha_J \,|\, \mu_A = \mu_B\,) = 1 - \alpha \tag{2.16}$$

and can be computed using recursive numerical integration.

Wang & Tsiatis defined the optimal boundaries as those that require the least number of patients for detecting a treatment difference at a given significance level α and power $1 - \beta$, and proposed a class of boundaries indexed by a single parameter Δ as follows:

$$|S_j| \geq \Gamma(\alpha, J, \Delta)\, j^{\Delta}, \qquad j = 1, ..., J \tag{2.17}$$

$\Gamma(\alpha, J, \Delta)$ is a positive constant derived to satisfy (2.16). Table 2.4 presents the maximum and expected sample size using Pocock's, O' Brien & Fleming's, and Wang & Tsiatis's method computed for optimal choice of Δ.

	$1-\beta$			
	0.80	0.90	0.95	0.99
$J = 1$				
	31.4 (31.4)	42.0 (42.0)	52.0 (52.0)	73.5 (73.5)
$J = 2$				
Pocock	26.8 (34.9)	32.6 (46.2)	37.3 (56.8)	46.3 (79.5)
O' Brien & Fleming	28.3 (31.6)	35.8 (42.3)	41.7 (52.3)	51.9 (73.9)
Wang & Tsiatis *(Optimal choice of Δ)*	26.7 (33.9)	32.6 (46.0)	37.3 (56.9)	46.3 (79.0)
$J = 3$				
Pocock	25.7 (36.6)	30.3 (48.4)	33.7 (59.2)	39.6 (82.5)
O' Brien & Fleming	26.9 (31.9)	33.6 (42.7)	39.0 (52.8)	49.1 (74.5)
Wang & Tsiatis *(Optimal choice of Δ)*	25.5 (34.6)	30.3 (47.9)	33.7 (60.3)	39.4 (85.5)

Table 2.4: Expected (maximum) sample size for 1, 2 and 3 maximum interim analyses J using Pocock's, O' Brien & Fleming's, and Wang & Tsiatis's (optimal boundaries) methods for a significance level of 0.05. [47]

2.7 Modified Wilcoxon statistic

Having a different approach, Slud & Wei (1982) [46] assumed that patients enrol sequentially to treatment arms A and B, following a superposition of two independent Poisson distributions with $\lambda_c(t) = c(\lambda_A + \lambda_B)$ $(c > 0)$ and random loss to follow-up.

Let $Z_k \in \{0, 1\}$ be an index variable denoting assignment to treatment arm A or B, and X_k and Y_k being the survival and loss-to-follow-up time of patient $k\,(k = 1, ..., N)$, respectively. The modified Wilcoxon statistic $W_c(t)$ is used for testing the null hypothesis of $F_A = F_B$ at each interim analysis, with F_A and F_B denoting the distribution function of the survival time. $W_c(t)$ is given by:

$$W_c(t) = \frac{1}{\sqrt{n_A(t)\, n_B(t)\, n(t)}} \sum_{k=1}^{n(t)} \sum_{l=1}^{n(t)} Z_k(1 - Z_l)\, \phi\left(T_k(t), \Delta_k(t); T_l(t), \Delta_l(t)\right) \quad (2.18)$$

with

$$\phi\left(T_k(t), \Delta_k(t); T_l(t), \Delta_l(t)\right) = \begin{cases} 1\,, & if \;\; T_k(t) < T_l(t) \;\; and \;\; \Delta_k(t) = 1 \\ -1\,, & if \;\; T_k(t) > T_l(t) \;\; and \;\; \Delta_l(t) = 1 \\ 0\,, & otherwise \end{cases}$$

denoting the Gehan-Gilbert score function, $T_k(t) = min\,(X_k, Y_k, t - t_k)$, $\Delta_k(t) = I[X_k \leq min\,(Y_k, t - t_k)]$, t_k the arrival time of patient k, and I the indicator function.

At time point t_j $(1 \leq j \leq J)$, the stopping boundary d_j can be derived as follows:

$$P\left(|V_1| < d_1\,, \ldots\,, |V_{j-1}| < d_{j-1}\,, |V_j| \geq d_j\right) = \alpha_j \quad (2.19)$$

with $(V_1, V_2, ..., V_j)$ a multivariate normal having $\boldsymbol{\mu} = \mathbf{0}$ and covariance:

$$\sigma_{im} = \frac{\hat{\sigma}(t_i, t_m)}{\sqrt{\hat{\sigma}(t_i, t_i)\, \hat{\sigma}(t_m, t_m)}} \;\; (1 \leq i \leq m \leq j) \quad (2.20)$$

while the significance level α_j must have been pre-defined as $\alpha = \sum_{j=1}^{J} \alpha_j$.

- If $|W_c(t_j)| \geq d_l \sqrt{\hat{\sigma}(t_j, t_j)}$ the trial is stopped and H_o is rejected.

P-value is defined as:

$$p_j = P\left(|V_1| < d_1\,, \ldots\,, |V_{j-1}| < d_{j-1}\,, |V_j| \geq W_c(t_j) \sqrt{\hat{\sigma}(t_j, t_j)}\right) \quad (2.21)$$

University of Piraeus

Table 2.5 presents stopping boundaries d_j and p-values for repeated two-sided tests at four time points $t_j \in \{3, 6, 9, 12\}$ for two different sequences of α_j. For further detail on this example refer to [46].

	t_j			
	3	6	9	12
α_j	0.0075	0.0125	0.015	0.015
boundary d_j	2.674	2.478	2.307	2.162
p-value	0.321	0.020	0.008	0.008
α_j	0.005	0.010	0.015	0.020
boundary d_j	2.807	2.560	2.325	2.095
p-value	0.321	0.021	0.003	0.009

Table 2.5: Two-sided boundary d_j and p-value at 4 interim analyses for two choices of significance levels α_j with overall $\alpha = 0.05$ using modified Wilcoxon test scores. Values correspond to data from a study on prostate cancer conducted by the Veterans Administration Cooperative Urological Research Group (VACURG) presented in [46].

2.8 The F-H-O procedure

Fleming, Harrington & O' Brien (1984) [49] used Slud & Wei's technique [46] (discussed in 2.7) to develop a design that preserves the appealing aspects of O' Brien & Fleming's method, whilst the maximum number of interim analyses may be increased in certain cases such as low event or accrual rate.

Assume that the joint distribution of the test statistic is a multivariate normal such that:

$$\{S_1, S_2, ..., S_J\} \sim \boldsymbol{N(0, \Sigma)} \tag{2.22}$$

and that the J statistics form an independent increment process, so that for every $j < k$ (with $j \in \{1, ..., J\}$):

$$E(S_j S_k) = var(S_j) \tag{2.23}$$

The probability that the trial will be terminated for efficacy under H_o at time t_j is given by:

$$\pi_j = P\left(\mid Z_1 \mid < c_1, ..., \mid Z_{j-1} \mid < c_{j-1}, \mid Z_j \mid \geq c_j \mid H_o\right) \tag{2.24}$$

with

$$\alpha = \pi_1 + \pi_2 + ... + \pi_{J-1} + \pi_J \tag{2.25}$$

Z_j the standard normal variate:

$$Z_j = \frac{S_j}{\sigma_j} \tag{2.26}$$

and c_j the stopping boundaries.

Slud & Wei [46] pointed out that if the joint distribution of $(Z_1, ..., Z_j)$ is known for every j, then a group sequential procedure can be formulated by specifying $(\pi_1, ..., \pi_j)$ and then recursively solving for $(c_1, ..., c_j)$ using (2.24). Note that α_j^* (the nominal level at which the j test is performed) satisfies the equality of:

$$c_j = z_{1-\alpha_j^*/2} \tag{2.27}$$

with z_α denoting the α-percentile of the standard normal distribution.

Assuming

$$\pi_1 = \pi_2 = ... = \pi_{J-1} = \pi \tag{2.28}$$

it follows that

$$\pi_J = \alpha - (J-1)\pi \tag{2.29}$$

Two ways exist to obtain the joint distribution of the test statistic, and subsequently to set the group sequential design−by assuming (2.22), (2.23) and either consistently estimate σ_j or by assuming that:

$$\sigma_j = j/J \tag{2.30}$$

(an assumption that often does hold).

Having specified α, J and π and assuming (2.28), stopping boundaries c_j can be determined by either recursive numerical multivariate integration or computer simulations (e.g. Monte Carlo simulation).

Table 2.6 gives the nominal levels α_j^* for $J = 2$ and $J = 3$ interim looks, for different values of $\pi\,(= \alpha_1^*)$ assuming (2.30). Values were obtained through numerical integration.

F-H-O method relates to an older method discussed by Haybittle [55] in 1971.

$J=2$		$J=3$		
1	2	1	2	3
0.005	0.048	0.0025	0.0030	0.0483
0.010	0.045	0.0050	0.0061	0.0459
0.015	0.042	0.0075	0.0094	0.0429
0.020	0.038	0.0100	0.0127	0.0395
0.025	0.034	0.0125	0.0161	0.0356

Table 2.6: Nominal significance levels using the F-H-O method for $J = 2$ and $J = 3$ looks, for overall α of 0.05. The nominal level of the first look (highlighted cells) is pre-selected. [49]

2.9 Discrete sequential boundaries

In 1986, Bauer [50] constructed conservative critical regions when neither the maximum number of interim analyses nor the the group sizes are pre-defined. Critical regions under H_o are determined by the following two conditions:

$$\left.\begin{aligned} P\left(Y_1 \in w_1\right) &\leq \alpha^*(n_1/n) \\ P\left(\left(\bigcap_{j=1}^{i-1}(Y_j \notin w_j)\right) \cap \left(Y_i \in w_i\right)\right) &\leq \alpha^*\left(\sum_{j=1}^{i} n_j/n - \sum_{j=1}^{i-1} n_j/n\right) \end{aligned}\right\} \tag{2.31}$$

where Y is the test statistic, α^* the Lan & DeMets's spending function and $i \in \{2, ..., K\}$ the interim analysis. By using (2.31) and the following boundary condition for Lan & DeMets's α:

$$\left.\begin{aligned} \alpha^*(t), \quad (0 \leq t \leq 1) \\ \alpha^*(1) = \alpha \end{aligned}\right\} \tag{2.32}$$

global α will be achieved if:

$$1 - P\left(\left(Y_1 \notin w_1\right) \cap ... \cap \left(Y_K \notin w_K\right)\right) \leq \alpha \tag{2.33}$$

From (2.31) it is clear that the only requirement of this approach is the knowledge of the joint distribution of the test statistics at the last two inspection times.

Unfortunately, establishing the joint distribution of test statistics for equal or unequal group sample sizes (n_i) can be a quite burdensome procedure. Bauer [50] allowed for dimension reduction of the left-hand side of (2.31) by proving that:

$$P\left(\left(\bigcap_{j=1}^{i-1}(Y_j \notin w_j)\right) \cap (Y_i \in w_i)\right) \leq P(Y_i \in w_i) - P\left(\bigcup_{j=m}^{i-1}((Y_j \in w_j) \cap (Y_i \in w_i))\right) \tag{2.34}$$

for $i \in \{2, ..., K\}$ and $m \in \{1, ..., i\}$.

- For $m = i$ the right-hand side of (2.34) equals $P(Y_i \in w_i)$, constituting a very conservative Bonferroni-type approximation.

- For $m = i - 1$, the right-hand side of (2.34) is now approximated by:

$$P(Y_i \in w_i) - P\left((Y_{i-1} \in w_{i-1}) \cap (Y_i \in w_i)\right) \tag{2.35}$$

The approximation is unlikely to be improved by using a $m < i - 1$. Table 2.7 presents exact and approximated (using Bauer's method [50]) stopping boundaries of 5 sequential interim analyses having equal sample sizes and following a Lan & DeMets's spending function. Note that the approximate values are clearly close to the exact ones.

	c_1	c_2	c_3	c_4	c_5
			$\alpha = 0.025$		
Exact	2.58	2.49	2.41	2.34	2.28
Approx.	2.58	2.49	2.43	2.37	2.33
			$\alpha = 0.050$		
Exact	2.33	2.22	2.12	2.03	1.96
Approx.	2.33	2.22	2.14	2.08	2.03

Table 2.7: Comparison of the exact (Lan & DeMets's [44]) and approximate (Bauer's [50]) one-sided boundaries for 5 groups with equal group sizes and $\alpha^*(t) = \alpha t$. More information on this example in [50].

2.10 Stochastic curtailment

Using the stochastic curtailment approach to terminate a trial for futility, the stopping boundary depends on a measure of the probability that H_o will be rejected at the final analysis. In essence, at each interim look, the biometrician projects ahead and computes the probability of H_o rejection at the end of the trial. Stochastic curtailment often requires the determination of frequentist conditional power and Bayesian predictive power. [36]

Conditional power is the frequentist conditional probability that at final look J,

University of Piraeus

the test statistic would exceed its boundary for declaring significance. [51] The conditional power at each look is computed based on the assumption that the test statistic at the final analysis J is a weighted average of the test statistic at interim analysis j and the information that would be accumulated between j and J analyses. This method requires the distribution of the yet unobserved information. To do so, the biometrician, at each interim look, has to assume some particular value for the parameter of interest θ. Common choices for θ include the current crude estimate $\hat{\theta}$, the hypothesized value of θ under H_o (especially when stopping for efficacy is viewed as most probable) or the hypothesized value of θ under H_1 (especially when stopping for futility is considered). Also, it is possible that θ estimates vary across interim analyses. [36]

Conditional power can be determined as [36]:

$$C_j\left(l_J^{(S)}, \theta_1\right) = P\left(S_J \leq l_J^{(S)} \,|\, S_j = s_j; \theta = \theta_1\right) \tag{2.36}$$

with $l_J^{(S)}$ denoting the efficacy boundary at the final look J, S the test statistic and θ_1 the value of θ under H_1.

The conditional power C_j serves as a test statistic with early stopping for futility being decided when:

$$C_j \geq d_j^{(C)} \qquad j \in (1, 2, ..., J-1) \tag{2.37}$$

In this case, $d_j^{(C)}$ is the stopping boundary for futility of the new treatment. It is a common approach that biometricians use a constant boundary $d_j^{(C)}$ across interim looks, with values of 0.1 or 0.2 being the most popular choices for $d_j^{(C)}$. However, as with other approaches, a boundary shape function can make early stopping more efficient. [36]

2.11 Assumption violations

In 1992, Proschan, Follmann & Waclawiw [16] examined the degree to which assumption violations can inflate type I error in several sequential designs. For simplicity reasons, authors assumed a one-sided alternative hypothesis (analogous results can be obtained for a two-sided test). Early termination can occur only due to strong evidence for H_o rejection.

Unequally spaced looks

A number of group sequential methods like those of Pocock's [43] and O' Brien & Fleming's [37], assume equally spaced information time between interim analyses.

Accidental violation of this assumption can be due to various factors such as non-constant accrual or event rate.

What happens if the information times $t_j = j/J$ are not equally spaced and for example, after three looks (J=3) and overall α=0.05, the actual global level is as high as 0.068? [16]

Pocock's and O' Brien & Fleming's methods

In O' Brien & Fleming's method, if J=5 and α=0.05, the worst-case scenario for the resulting α is 0.078. And if one chooses to look too many times close to information time 0, the probability of falsely rejecting H_o will be close to 1. Not surprisingly, this problem is largely overcome if the biometrician agrees not to look at the data before some information time δ. By using for example Pocock's method, when $\alpha = 0.05$, $J = 3$ and $\delta = 0.30$, the upper bound of α is as low as 0.057. [16]

Tables 2.8 & 2.9 present 2 scenarios (A & B) with 4 designs each using Pocock's and O' Brien & Fleming's methods, respectively. Exact type I error rates were calculated using recursive numerical integration. Note that for Pocock's method, the α inflation is subtle when the first look is done as scheduled (Scenario A; Table 2.8), whilst the inflation is moderate if the first look is done 50% earlier than expected (Scenario B; Table 2.8). On the other hand, O' Brien & Fleming's method is more sensitive to unequally spaced looks as α inflation is sharp even when the first look is done as scheduled (Table 2.9).

Scenario	t_1	t_2	t_3	t_4	t_5	$\alpha = 0.025$	$\alpha = 0.05$
A	0.50	1.00				0.0250	0.0500
	0.33	0.60	1.00			0.0251	0.0503
	0.25	0.40	0.70	1.00		0.0253	0.0507
	0.20	0.28	0.44	0.68	1.00	0.0256	0.0511
B	0.25	1.00				0.0271	0.0547
	0.17	0.42	1.00			0.0277	0.0558
	0.13	0.21	0.48	1.00		0.0281	0.0564
	0.10	0.19	0.37	0.64	1.00	0.0285	0.0569

Table 2.8: Type I error inflation with Pocock's method [43] for two values of α (t denoting the information time). Scenario A: First and last looks as scheduled, intermediate looks chosen to inflate global α. Scenario B: First look at 50% earlier than scheduled, last look as scheduled, intermediate looks chosen to inflate global α. For more information on the evaluation method see [16].

University of Piraeus

Scenario	t_1	t_2	t_3	t_4	t_5	$\alpha = 0.025$	$\alpha = 0.05$
A	0.50	1.00				0.0250	0.0500
	0.33	0.40	1.00			0.0269	0.0542
	0.25	0.33	0.40	1.00		0.0281	0.0564
	0.20	0.28	0.36	0.44	1.00	0.0289	0.0575
B	0.25	1.00				0.0257	0.0524
	0.17	0.25	1.00			0.0279	0.0565
	0.13	0.21	0.30	1.00		0.0292	0.0590
	0.10	0.19	0.28	0.37	1.00	0.0299	0.0600

Table 2.9: Same as Table 2.8 for O' Brien & Fleming's method [37]. [16]

Planning based on data trends

When using methods that do not require to specify in advance the number of inspection times (e.g. an α-spending function or the F-H-O approach [49]), biometricians are often tempted to determine the number of future looks based on the accumulated data, even if this act could inflate type I error. In this case, an assumption violation (which is not accidental, though quite natural) may occur. [16]

[1] *The authors make it clear that look times should not be chosen based on trends in the data, but how much α inflation will occur if this is done?* [16]

The Slud & Wei / F-H-O procedure

In the Slud & Wei [46] / F-H-O procedure [49], H_o is rejected at information time t_j, if:

$$Z(t_j) > C(t_1, ..., t_j) \tag{2.38}$$

with $C(t_1, ..., t_j)$ chosen so that:

$$P\left(Z(t_1) < C(t_1), ..., Z(t_{j-1}) < C(t_1, ..., t_{j-1}), Z(t_j) > C(t_1, ..., t_j)\right) = \pi_j \tag{2.39}$$

with

$$\sum_{j=1}^{J} \pi_j = \alpha \tag{2.40}$$

Using the F-H-O method, the answer to [1] is *"not much"* since $(\pi_1, \pi_2, ..., \pi_{J-1})$ are quite small compared to π_J. However:

If π_J is not the bulk of α, we may grossly inflate α by basing future look times on data trends. [16]

Proschan et al. (1992) [16] gave an example: Suppose that $\pi_{J-1} \simeq \alpha$ and the other

π_j close to 0. One can look very close to information time 0, say time ϵ_1, using virtually a negligible part of α, and therefore having no chance of rejecting−he simply notes whether $Z(\epsilon_1) > z_\alpha$. If so, he chooses to spend $(\pi_2, ..., \pi_{J-1})$ almost immediately thereafter, with rejection taking place almost certainly when π_{J-1} is spent. If $Z(\epsilon_1) < z_\alpha$, the biometrician may be tempted to look at information time ϵ_2 again close to information time 0 but such that ϵ_1/ϵ_2 is very small (to ensure that $Z(\epsilon_1)$ and $Z(\epsilon_2)$ are independent). If he continues in this vein and at least one $Z(\epsilon_i)$ exceeds z_α, rejection is almost guaranteed. The probability of rejection is approximately:

$$P\left(\bigcup_{i\leq J-1} Z(t_i) > z_\alpha\right) \approx 1 - (1-\alpha)^{J-1} \tag{2.41}$$

Note that, according to (2.41), as J increases, the probability of rejection increases too.

When J is large, the probability of rejecting H_o is almost 1! This procedure can be open to extreme abuse depending on the choice of π_j. [16]

Table 2.10 presents type I error inflation when using less *"abusive"* ways of scheduling interim analyses based on data trends. For instance, when $\pi_j = \alpha/J$ (Scenario A), or when $\pi_J = \alpha/2$ and $(\pi_1, \pi_2, ..., \pi_{J-1}) = \alpha/[2(J-1)]$ (Scenario B). At information time $1/J$, if $Z(1/J)$ exceeds its boundary, or if it does not but is close enough so that H_o will be almost certainly rejected, $(\pi_2, ..., \pi_{J-1})$ are subsequently spent almost immediately. In another case, trial continues to look $2/J$ and so on. Note that in both cases there is a substantial α inflation, even in scenario B suggested by Fleming, Harrington & O' Brien (1984) [49]. As expected, type I error increases with the maximum number of looks J.

J	Scenario A	Scenario B
2	0.050	0.050
3	0.062	0.059
4	0.075	0.066
5	0.087	0.072

Table 2.10: Limiting case of "unintentional abuse": Type I error inflation with Slud & Wei [46] / F-H-O procedure [49] ($\alpha = 0.05$). Scenario A: $\pi_j = \alpha/J$. Scenario B: $\pi_J = \alpha/2$, $(\pi_1, \pi_2, ..., \pi_{J-1}) = \alpha/[2(J-1)]$. For further information see [16].

Alpha-spending functions

A substantial difference between Slud & Wei / F-H-O procedure and the use of an α-spending function is that, as pointed out in paragraph 2.3, in the latter case, the

amount of α to be spent at the next look is integrally dependent on its information time (see (2.8) for α-spending functions, and (2.38) & (2.24) for the Slud & Wei / F-H-O procedure). [16]

If one looks the data at time t, then unlikely the F-H-O method, he will pay a price for looking again soon after t, say $t+\epsilon$. The reason is that he will be able to spend only the $\alpha(t+\epsilon)-\alpha(t)$ amount of α and therefore the constant necessary to conclude significance will be large.

The use of an α-spending function provides some protection if $\alpha(t)$ is continuous against basing future look times on current data trends, though again the authors make it clear that this is prohibited. [16]

Table 2.11 presents a worst-case scenario of α inflation with 3 common use α-spending functions [44] and 3 looks at 10%, 25% and 50% of information time. Note that type I error inflation is mild in all cases. For instance, using $\alpha_1(t)$ and having $\alpha = 0.05$, when one abusively looks at the data at $t = 0.10$, one will get an actual type I error of approximately 0.053 which is only 6% larger than the required one.

	$\alpha = 0.025$			$\alpha = 0.050$		
	$\delta_1 = 0.10$	$\delta_2 = 0.25$	$\delta_3 = 0.50$	$\delta_1 = 0.10$	$\delta_2 = 0.25$	$\delta_3 = 0.50$
$\alpha_1(t)$	0.0261	0.0256	0.0260	0.0528	0.0514	0.0520
$\alpha_2(t)$	0.0272	0.0269	0.0260	0.0537	0.0531	0.0517
$\alpha_3(t)$	0.0268	0.0267	0.0261	0.0529	0.0529	0.0518

Table 2.11: Worst-case α inflation with 3 common used α-spending functions (3 looks at 10%, 25% and 50% of information time) for two different values of global α. $\alpha_1(t) = 2 - 2\Phi(z_{\alpha/2}/\sqrt{t})$; $\alpha_2(t) = \alpha\, log(1 + (e-1)t)$; $\alpha_3(t) = \alpha t$. [44] Values were obtained through numerical integration (this method of abuse is computationally intensive even with only 3 looks−most values required many hours of CPU time on a micro Vax II to be obtained). For further information see [16].

2.12 Further sources

For more information on the group sequential design, the reader may refer to review articles and books by Jennison & Turnbull (1991, 1999) [40] [56], Turnbull (1997) [57] and Whitehead (1997) [58]. For the adaptive design, interesting reviews are those of Wassmer's (2000) [41] and Maca's (2006) [59].

Multi-stage design

3.1 Introduction

In clinical research, there are often several therapeutic regimens - candidates to be evaluated against a control (a standard treatment). These regimens may refer to different treatments, different doses or defining the most efficient duration of exposure to the new intervention in terms of risk-benefit ratio. Sometimes selection is made in an informal manner based on non-randomised pilot studies often from various sources. [60] This renders treatment selection a poorly-controlled and subtly arbitrary process. Another approach includes selecting an appropriate treatment or therapeutic dose in a phase II trial and then compare it with the standard of care in a phase III trial. [61] However, if selection and comparison phases could be combined in a single trial, sample size will be reduced, and a great deal of time and cost will be saved. Table 3.1 (an example by Todd & Stallard [62]) illustrates the magnitude of sample size saving using either a two-stage or a seamless design (the latter will be discussed in chapter 4).

Treatment selection is typically performed at the first interim analysis in where multiple treatments are compared with a control (exploratory phase), but only one (or some) continues along with it to the second stage (confirmatory phase), or it may be done by dropping arms in a series of interim analyses (multi-stage scheme). In the two-arm scheme, obtaining a significant difference at an interim analysis implies trial termination, and therefore the stopping boundaries used to close an arm are identical to those used for closing the entire trial. However, this is not the case for multi-armed trials. In the pairwise comparison scheme, several treatments are usually compared with a control and arms are gradually rejected. At each interim analysis, several test statistics are compared with the corresponding critical values and each interim analysis has to answer two questions: the number of treatments which will continue along with the control, and whether the entire trial must be terminated. The latter may happen after the last treatment arm has been proven superior or inferior to the control, or when all planned patients have been evaluated. However, in practice

other factors may also be taken into consideration when deciding whether to drop or keep a treatment arm. For instance, risk-benefit analysis may allow inferior arms with low rate of serious adverse events to continue and reject superior arms with higher toxicity rate.

A classic approach demands that the experimental arm with the largest effect estimate being selected at the first interim analysis, whilst its test statistic ξ_1 being compared with two numeric constants (stopping boundaries) C_{11} and C_{12}, so that when $\xi_1 < C_{11}$ the trial is stopped for futility, whilst when $\xi_1 > C_{12}$ the trial is stopped for efficacy, with $C_{11}, C_{12} \in (-\infty, +\infty)$. In any other case, the experimental treatment continues along with the control to the confirmatory phase, in where data from both two stages are looked at a series of interim analyses. At interim analysis j, if $\xi_j > C_{j2}$ the trial is stopped and H_o is rejected, whilst if $\xi_j < C_{j1}$ the trial is stopped for futility. If $\xi_j \in [C_{j1}, C_{j2}]$ the trial continues to the $j+1$ interim analysis. At the final analysis J, if $\xi_J > C_J$ the trial is stopped with H_o rejection, whilst in another case, the experimental treatment cannot be concluded to be effective. Stopping boundaries are chosen in such a way to maintain an overall type I error, whilst sample size is calculated based on parameter estimates to fulfil a power prerequisite under the alternative hypothesis.

In multi-armed comparative trials, multiplicity arises from two sources: multiple comparisons of different arms and repeated significance testing that is done in the context of a group sequential design. [63] Note than when multiple treatments are compared with a control, a type I error can occur in several ways, whilst power slipping away from its traditional definition, now can be expressed as a function of both type II and type I error. [64][62]

Early work on multiple decision, or otherwise known as ranking and selection process, started in the mid-1950s by Bechhofer [65] who considered the normal means problem under what today is known as the indifference zone approach (see Appendix A), and by Paulson (1964) [20] who assumed normal distributions with equal variances for selecting the population with the largest mean in the context of a sequential elimination procedure. Hoel & Mazumdar (1968) [66] extended Paulson's elimination process to other distributions. Bechhofer et al. (1968) [22] proposed the Koopman-Darmois exponential distribution family. Hsu & Edwards (1983) [26] proposed a method to select the best treatment(s) based on confidence sets. In 1994, Follman et al. [67] modified some broadly used α-maintaining methods to allow for pairwise comparisons in a multi-arm setting demanding equal evidence against all pairwise comparisons. Maximum number of interim analyses does not have to be specified in advance. Unfortunately, this approach increases expected sample size and renders it suitable only for large-scale clinical trials. [60]

In 1990, a two-stage approach by Schaid et al. [68] for time-to-event endpoints used the logrank statistic for comparing between treatments assuming risk proportionality. Stallard & Todd (2003) [61] generalised the designs of Thall et al. [60] for binary data and Schaid et al. for failure times in a simple two-stage sequential design suitable for binary, normally distributed and failure time response. Treatment selection is based on comparing test statistics among investigational treatments, with the best arm compared against the control in the context of a formal hypothesis testing. Bischoff & Miller (2005) on the other hand, performed treatment selection by comparing mean estimates [64].

The mathematical background of some of the aforementioned approaches is presented in this chapter. Emphasis is given in the sequential design using continuous endpoints (including failure-time data) to measure response.

Design	**Total sample size**
Four-arm phase III	4,000
Four-arm phase II followed by a two-arm phase III (fixed sample)	2,100
Four-arm phase II followed by a two-arm phase III (group sequential)	1,440 *(under H_1)*
Combined phase II/III	1,390 *(under H_1)*

Table 3.1: An example by Todd & Stallard: Expected total sample size for different designs. For more information see [62].

3.2 Pairwise comparisons

Lan & DeMets's method [44] for the two arm group sequential scheme dictates that H_o shall be rejected at time t_j if $|Z(t_j)| > c(t_1, ..., t_j)$ with stopping boundary $c(t_1, ..., t_j)$ being chosen so that:

$$P(reject\ H_o\ at\ t_1\ \cup\ ...\ \cup\ reject\ H_o\ at\ t_j) = \alpha^*(t_j) \tag{3.1}$$

with $\alpha^*(t)$ denoting the α-spending function. Follman et al. (1994) [67] adjusted (3.1) to account for pairwise comparisons in a multi-armed scheme. Now the condition is expressed as:

$$P(\cup_{ik \in I}(reject\ H_{ik}\ at\ t_1\ \cup\ ...\ \cup\ reject\ H_{ik}\ at\ t_j)) = \alpha^*(t_j) \tag{3.2}$$

In other words, stopping boundaries shall be chosen so that the cumulative chance of any H_{ik} rejection up to or at interim analysis t_j to be equal to $\alpha^*(t_j)$, with H_{ik}

University of Piraeus

denoting the null hypothesis for the equality of treatments i and k, and I the set of hypotheses under testing. H_{ik} is rejected at time t_j if $|Z_{ik}(t_j)| > c(I, t_1, ..., t_j)$. Note that the same boundary c is used for all pairwise testing at a given interim analysis.

Follman et al. also suggested a Pocock's [43] and an O' Brien & Fleming's [37] analogue for hypotheses testing. Having a set of I comparisons and J interim looks, Pocock's method translates as:

$$P(|Z_{ik}(j/J)| > C_P(I, J) \; for \; some \; (i, k) \in I \; and \; some \; j = 1, ..., J) = \alpha \qquad (3.3)$$

whilst O' Brien & Fleming's method translates as (see also Table 3.2):

$$P(|Z_{ik}(j/J)| > C_{OF}(I, J)\sqrt{\frac{J}{j}} \; for \; some \; (i, k) \in I \; and \; some \; j = 1, ..., J) = \alpha \qquad (3.4)$$

Pairwise comparisons approach usually requires a large sample size to reject H_o. Alternatives to this method will be discussed in this chapter.

		Treatments vs Control		All pairwise comparisons	
No. of arms	J	Pocock	O' Brien	Pocock	O' Brien
2	1	1.96	1.96	1.96	1.96
	2	2.18	1.98	2.18	1.98
	3	2.29	2.00	2.29	2.00
3	1	2.24	2.24	2.39	2.39
	2	2.45	2.25	2.60	2.40
	3	2.56	2.27	2.70	2.42
4	1	2.39	2.39	2.64	2.64
	2	2.60	2.40	2.83	2.64
	3	2.70	2.42	2.93	2.66

Table 3.2: Critical values for a multi-arm analogue to Pocock's and O' Brien & Fleming's methods having J interim analyses using Bonferroni approach at an overall two-sided significance level of $\alpha = 0.05$ (equal variances assumed). [67]

3.3 Logrank statistic in a two-stage scheme for failure times

Schaid et al. (1990) [68] used logrank statistic for comparing between treatment arms assuming proportional hazard rates $\lambda_o(t)$ and $\lambda_k(t)$ for patients of different treatment arms. Having n_1 patients at each of the treatment $K + 1$ groups (K investigational arms plus the control) at the end of the exploratory phase $t = t_1$,

accrual is terminated at time t_1 in the following two cases:

- if all the K statistics are lower than a C_1 boundary (*stopping for futility*)
- if at least one statistic exceeds an upper boundary C_2 (*stopping for efficacy*)

In any other case, accrual is continued for the control group and all experimental groups for which $C_1 \leq T^k(t_1) \leq C_2$, with $T^k(t)$ denoting the logrank statistic for comparing a k experimental treatment ($k = 1, ..., K$) with the control at time point t. At the end of confirmatory phase, successful treatments should have $T^k(t_2) > C_3$.

By using the Bonferroni approximation for the overall significance level, under H_o, the expected total sample size is given by:

$$N = (K+1)n_1 p_0 + \sum_{j=1}^{K} [n_2(j+1) + n_1(K-j)]\, p_j \tag{3.5}$$

with j denoting the comparison number ($j = 1, ..., K$), p_0 the probability of accrual termination at the end of the exploratory phase and p_j the probability that accrual of j experimental treatments (along with the control) will continue to the next phase (for p_0 and p_j evaluation see [68]).

For a fixed type I error (α) and power ($1-\beta$), an optimal design is the one with the smallest expected sample size under H_o. For simplicity, Schaid et al. assumed a uniform accrual rate in $[0, t_\alpha]$, an exponential survival time distribution with parameter $\lambda_o(t) = \lambda_o$, and no loss to follow-up (other forms of accrual, survival and censoring may be assumed as well). The parameters that need to be pre-specified are α, $1-\beta$, K, λ_o, θ, follow-up time $t_2 - t_\alpha$ and accrual rate, with θ denoting the hazards ratio $\lambda_o(t)/\lambda_k(t)$, whilst the parameters to be calculated are the stopping boundaries C_1, C_2, C_3 and the sample size of each stage n_1, n_2. C_1 and C_3 boundaries were set equal to:

$$\left.\begin{aligned} C_1 &= \frac{1}{2}\sqrt{d_1}\, log\tilde{\theta} \\ C_3 &= \Phi^{-1}(1-\alpha) \end{aligned}\right\} \tag{3.6}$$

with d_1 denoting the expected number of deaths at the end of the exploratory phase, and $\tilde{\theta}$ the minimum hazards ratio that need to be recorded before continuing accrual.

Using (3.5) and having defined (α, $1-\beta$, K, λ_o, θ, C_1 and C_3), the optimal choice of (n_1, n_2, C_2) is the one leading the smallest expected total sample size N under H_o. Table 3.3 presents the expected sample size (asymptotically derived and simulated), power and type I error for two different combinations of accrual and hazard rate (c). This design can offer a substantial saving of sample size when the hazard rate of the

experimental treatment is large relative to the accrual rate. [68]

			$c = 50$				$c = 100$			
θ	K	$1-\beta$	$1-\beta$	n	n	α	$1-\beta$	n	n	α
		Asym.	*Sim.*	*Asym.*	*Sim.*	*Sim.*	*Sim.*	*Asym.*	*Sim.*	*Sim.*
1.5	2	0.80	0.81	124.9	122.2	0.043	0.80	142.0	140.7	0.050
		0.90	0.90	160.8	161.4	0.040	0.87	179.9	180.1	0.048
	4	0.80	0.77	111.7	112.0	0.044	0.80	121.8	123.5	0.049
		0.90	0.90	141.9	142.2	0.048	0.88	152.7	151.0	0.047
2.0	2	0.80	0.81	56.0	57.6	0.054	0.80	67.2	67.0	0.049
		0.90	0.89	70.7	70.0	0.058	0.89	84.2	84.2	0.065
	4	0.80	0.76	46.9	47.0	0.044	0.79	54.8	54.6	0.055
		0.90	0.87	58.6	58.3	0.037	0.89	67.8	68.1	0.046

Table 3.3: Expected sample size per treatment arm (n) (asymptotically derived and simulated), Bonferroni approximated overall type I error (α), and pairwise power ($1-\beta$) (asymptotically derived and simulated) for two values of accrual rate/hazard rate (c) (nominal overall type I error set to 0.05). [68]

3.4 Other sequential two-stage designs

More recently, Stallard & Todd [61] proposed a two-stage design where selection is made from a small number of treatment arms, and subsequently the most efficient treatment is evaluated. At the first interim analysis, assuming equal variances among the experimental arms, the treatment which displayed the largest observed score is compared with the control. If a treatment difference is proven, then the trial is terminated, otherwise the best treatment continues along with the control to stage II. Authors adopted a number of settings for the parameter of interest θ and the score statistics Z given by Whitehead [58] for binary and normally distributed data with stopping boundaries depending on the observed Fisher's information table.

Stallard and Todd's design presents a series of benefits against other designs, including its simplicity, its applicability to various types of endpoints and the ability to incorporate covariate information. Note that treatment selection is not done by hypothesis testing, rather it is based on comparing test statistics of the investigational arms assuming homoscedasticity.

Bischoff & Miller (2005) [64] published an adaptive two-stage design for choosing the best treatment among a set of three different arms (including the control arm). If Y_{ijk} is the response of the ith patient to treatment j at stage k with $k = 1$ denoting

the first stage and $k = 2$ the second, then Y_{ijk} can be viewed as:

$$Y_{ijk} = \mu_j + \epsilon_{ijk} \tag{3.7}$$

independently for each i, j and k, with $j = 0$ denoting the control arm. In a superiority design, the hypotheses testing writes as follows:

$$\begin{aligned} H_o &: max(\mu_1, \mu_2) \leq \mu_o \\ H_1 &: max(\mu_1, \mu_2) > \mu_o \end{aligned} \tag{3.8}$$

with type I error given by:

$$\alpha = P(H_o\, rejected, \hat{\theta} = 1)\, I[\mu_1 \leq \mu_o] \;+\; P(H_o\, rejected, \hat{\theta} = 2)\, I[\mu_2 \leq \mu_o] \tag{3.9}$$

with θ denoting the best treatment, and I the indicator function. Note that, in contrast to traditional hypotheses testing, a type I error can occur even if the alternative is true. And power π now is related to both type II and type I error:

$$\pi = 1 - (\alpha + \beta)\, I[\mu_1 > \mu_o \; or \; \mu_2 > \mu_o] \tag{3.10}$$

with β denoting type II error ($\beta = P(H_o\, not\, rejected)\, I[\mu_1 > \mu_o \; or \; \mu_2 > \mu_o]$). Mean μ_j can be estimated by:

$$\overline{Y}_{.j1} = \frac{1}{n_1} \sum_{i=1}^{n_1} Y_{ij1} \qquad (j = 0, 1, 2) \tag{3.11}$$

with n_1 the number of allocated patients to each arm in stage I of the trial.

Treatment selection is done by comparing μ estimates of arms $j = 1$ and 2 so that $\overline{Y}_{.\hat{\theta}1} = max(\overline{Y}_{.11}, \overline{Y}_{.21})$, with test statistics expressed as:

$$\xi_1 = (\overline{Y}_{.\hat{\theta}1} - \overline{Y}_{.01}) \sqrt{\frac{n_1}{2S_1^2}} \tag{3.12}$$

with

$$S_1^2 = \frac{1}{3n_1 - 3} \sum_{j=0}^{2} \sum_{i=1}^{n_1} (Y_{ij1} - \overline{Y}_{.j1})^2 \tag{3.13}$$

denoting the variance estimate.

At interim analysis, stopping rules write as follows:

- if $\xi_1 > c$ stop and reject H_o
- if $\xi_1 < b$ $(b \leq c)$ stop and do not reject H_o
- if $\xi_1 \in [b, c]$ continue and compute N_2

University of Piraeus

with $c, b \in (-\infty, +\infty)$ and N_2 being equal to:

$$N_2 = N_2(S_1^2) = vS_1^2 - n_1 \tag{3.14}$$

and v an arbitrarily positive constant.

If trial continues to the second stage, another n_2 patients will be added to each of the two arms (the best treatment and the control), and final response is calculated based on data from patients of both first and second stage so that:

$$\hat{\mu}_j = \overline{Y}_{.j.} = \frac{1}{n_1 + n_2}(n_1\overline{Y}_{.j1} + n_2\overline{Y}_{.j2}) \qquad j \in \{0, \hat{\theta}\} \tag{3.15}$$

with test statistics now being equal to:

$$\xi_2 = (\overline{Y}_{.\hat{\theta}.} - \overline{Y}_{.0.})\sqrt{\frac{n_1 + n_2}{2S_1^2}} \tag{3.16}$$

Here H_o is rejected if $\xi_2 > u$ with u being chosen suitably.

Figure 3.1 presents the expected power as a function of μ_1 ($\mu_1 \in [0, \Delta]$) for $\mu_o = 0$ and $\mu_2 = \Delta$ with $max(\mu_1 - \mu_o, \mu_2 - \mu_o) \geq \Delta$. Table 3.4 presents three combinations of the values (n_1, b, c, u *and* v) and the expected total sample size (N), whilst table 3.5 shows the expected sample size for various (b, c) combinations. The method can be generalised to have more than three arms.

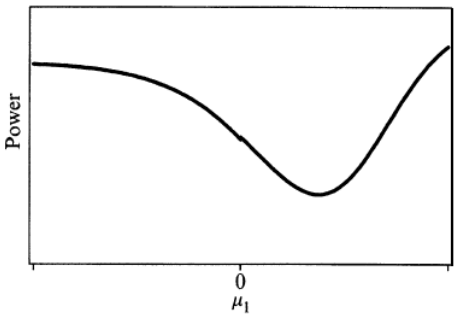


Figure 3.1: Typical shape of the power as a function of μ_1 for $\mu_o = 0$ and $\mu_2 = \Delta$, for fixed σ^2, n_1, b, c, u and v for a three-arm two-stage trial. [64]

n_1	b	c	u	v	N
30	-1.0	2.7	2.29	1.77	197
40	-0.6	2.7	2.27	1.58	191
50	-0.1	2.6	2.30	1.52	198

Table 3.4: Expected total sample size for three combinations of (n_1, b, c, u and v) for a three-arm, two-stage trial following the group sequential design. More information on this example in [64].

b	$c = 2.5$	$c = 2.6$	$c = 2.7$	$c = 2.8$	$c = 2.9$
-0.8	193.2	191.2	190.9	191.8	193.9
-0.6	193.0	191.0	190.7	191.7	193.9
-0.4	193.0	191.3	191.2	192.0	194.6

Table 3.5: Expected total sample size for different (b, c) combinations, having $n_1 = 38$ for a three-arm, two-stage trial following the group sequential design. More information on this example in [64].

University of Piraeus

Other designs

4.1 Introduction

The seamless phase II/III design

Seamless phase II/III trials have become quite popular due to their advantages upon other designs. An adaptive seamless II/III design combines the exploratory and the confirmatory stage in a single confirmatory trial allowing data collected from both phases to be used for the final efficacy assessment, thus leading to sample size reduction. Another advantage of such a design is the short span of time needed to seamlessly move from one phase to another.

Chow & Tu [69] named 4 categories of seamless designs according to study objectives and measuring endpoints of each stage: (a) same objectives/same endpoints; (b) same objectives/different endpoints; (c) different objectives/same endpoints; (d) different objectives/different endpoints. Note that by the term *"different objectives"*, authors mostly refer to designs described in chapter 3 (e.g. treatment selection in the exploratory phase and formal comparison in the confirmatory phase). In this chapter, focus is given on (a) and (b) designs.

To seamlessly move from a phase II to a phase III trial, patients must be treated under the same protocol. [62] However, an adaptive design allows for real-time protocol modifications which may refer to either the trial procedures (e.g. eligibility criteria, study dose, treatment duration, a change of endpoint, laboratory testing procedures, diagnostic procedures) or the statistical methods (e.g. sample size, randomization process, statistical design, a change of hypotheses to be tested). [69] These modifications must by planned in such a way to protect the validity and integrity of the trial. [70]

In 1994, Bauer & Köhne [42] proposed and evaluated a general method for multiple testing in the context of an adaptive two-stage scheme. The method is based on p-values obtained from the disjoint sample before and after the interim look. Efficacy analysis is then performed by combining the two p-values into a global test

University of Piraeus

statistic. This process is known as a p-value combination test. Authors used the Fisher's combination test [71] and assumed that the two p-values are independent and uniformly distributed under H_o. Figure 4.1 presents the rejection region for such a design.

More recently, Stallard [53] utilised the recursive numerical integration method [13] to compute the test statistics in a seamless design having multiple treatment arms. Although in a general treatment selection design, the number of treatments which continue along with the control in the next phase is determined by the data, in Stallard's design this is specified in advance, unless the entire trial is terminated early, either for efficacy or futility reason.

Multiple endpoints

Typically in clinical research, a single, clinically relevant endpoint is used to measure response to study treatment. Sample size determination, number and timing of interim looks and early termination rules are all related to the primary endpoint. However, the effect of a treatment is almost always multidimensional, rendering a single endpoint inadequate to present the full picture of benefit (including clinical benefit). Most clinical trials address different aspects of treatment benefit with secondary and exploratory endpoints. It is also quite usual, especially in the field of oncology, to conduct a translational study after the end of the trial to explore how study treatment could benefit different subgroups of patients.

The objective of a confirmatory trial is always addressed through its primary endpoint. Other endpoints that may be measured have only exploratory nature and no conclusion shall be drawn regarding them. The reason for that is that the entire statistical design (sample size, group size and interim analyses) is based on the primary endpoint. *So what if, to objectively measure clinical benefit of a new treatment, two endpoints must be used?*

In clinical research, it is not uncommon to use two co-primary endpoints to measure response. An efficacy endpoint and the toxicity rate, or two efficacy endpoints may constitute two primary endpoints of a clinical trial. Their responses can be correlated, and the degree of correlation is often non-foreseeable. In 2016, EMA released draft guidelines on multiplicity issues (CHMP, 2017), and in 2017, FDA issued guidance on multiple endpoints in clinical trials (FDA, 2017). These guidelines address the challenges raised by using multiple primary endpoints, including guidance on proper control of type I & II error.

In this Chapter, focus is given on the adaptive seamless II/III design and designs utilising more than one primary endpoint. Tables, figures and flowcharts are provided along with methods as reported by the authors.

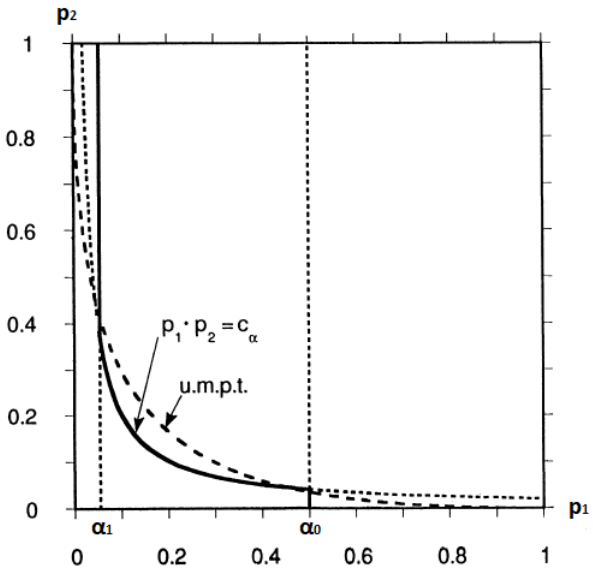


Figure 4.1: The p-value combination test. Rejection regions of the different test procedures in terms of the observed error probabilities p_1 and p_2 in the two stages of the trial. The uniformly most powerful test (u.m.p.t.) is based on the assumption of normally distributed test statistics with known variance and equal sample sizes at both stages. [42]

University of Piraeus

4.2 The p-value combination test approach

Bauer & Kieser (1999) [72] assumed a seamless design with treatment selection. At first interim analysis, the null hypothesis to be tested writes as follows:

$$H_o = H_{oA} = H_{o1} \cap H_{o2} \cap ... \cap H_{o(K-1)} \cap H_{oK} \tag{4.1}$$

with

$$\begin{aligned} H_{ok} &: \ \mu_o \geq \mu_k \\ H_{1k} &: \ \mu_o < \mu_k \end{aligned} \tag{4.2}$$

with $k = 1, ..., K$ and K denoting the number of experimental arms to be compared with the control.

H_{oA} is tested on significant levels $(\alpha_{AL}, \alpha_{AU})$ (stopping for futility or efficacy, respectively) providing a p-value p_A. Selected treatments after protocol adaptation continue to the second phase along with the control. Now the null hypothesis writes as follows:

$$H_{oB} = \bigcap_{l \in L} H_{ol} \tag{4.3}$$

with L a subset of K arms (note that by definition $H_{oA} \cap H_{oB} = H_{oA} = H_o$). A p-value p_B is obtained by testing H_{oB} with data collected only from patients joined in the second stage (disjoint sample).

Final decision for the rejection or not of the global null hypothesis H_o, using data from patients recruited in both phases, is performed by a combination test for p_A and p_B by assuming that p_A and p_B are independent and uniformly distributed under H_o (Fisher's combination test). The global level α is set equal to:

$$\alpha = \alpha_{AL} + \int_{\alpha_{AL}}^{\alpha_{AU}} \int_0^{c_B/p_A} dp_B dp_A = \alpha_{AL} + c_B \left[log(\alpha_{AU}) - log(\alpha_{AL}) \right] \tag{4.4}$$

with

$$c_B = exp\left[-\frac{1}{2} \chi_4^2 (1 - \alpha_B) \right] \tag{4.5}$$

Stopping rules write as follows:

Stage I:

- if $p_A \leq \alpha_{AL}$, the trial is stopped and H_o is rejected
- if $p_A \geq \alpha_{AU}$, the trial is stopped and H_o is not rejected
- if $\alpha_{AL} < p_A < \alpha_{AU}$, the trial continues to the second stage

Stage II:

- if $p_A p_B \leq c_B$, H_o is rejected
- if $p_A p_B > c_B$, H_o is not rejected

Fig. 4.2 presents a more general seamless sequential design using Fisher's combination test. For $\alpha_{AL} = c_B$ $(with\, \alpha = \alpha_B)$ the design corresponds to a non-stochastic curtailment, whilst for $(\alpha_{AU} \to \alpha_{AL}) \Rightarrow (\alpha_{AU} \to \alpha)$ the design corresponds to a fixed sample size test using only the first phase. Another design by the same authors—an adaptive seamless two-stage design for multiple inference of the treatment-control comparisons—is presented in Appendix B. [72]

4.3 Recursive numerical integration in a seamless design

In 2011, Stallard [53] used the method of recursive numerical integration (described in Chapter 2) in a seamless phase II/III design having multiple experimental treatments to be compared with a control. More specifically, he utilised two designs: a design proposed in 2003 by Stallard & Todd [61] in which only a single experimental arm moves to the second phase; and another design proposed by Stallard & Fiede (2008) [73] in which a pre-specified number of treatments continue along with the control.

In the first stage of Stallard's [53] design, the family of null hypotheses to be tested writes as follows:

$$H_{oi} : \theta_i \leq 0 \qquad i \in \{1, 2, ..., k\} \tag{4.6}$$

with θ_i denoting the effect of the T_i treatment relative to the control.

To control the familywise type I error in a *strong sense*, a one-sided test requires that:

$$P(reject\ any\ true\ H_{oi}) \leq \alpha \tag{4.7}$$

while a less stringent approach demands that:

$$P(reject\ any\ H_{oi} \mid H_{o1}, H_{o2}, ..., H_{ok}) \leq \alpha \tag{4.8}$$

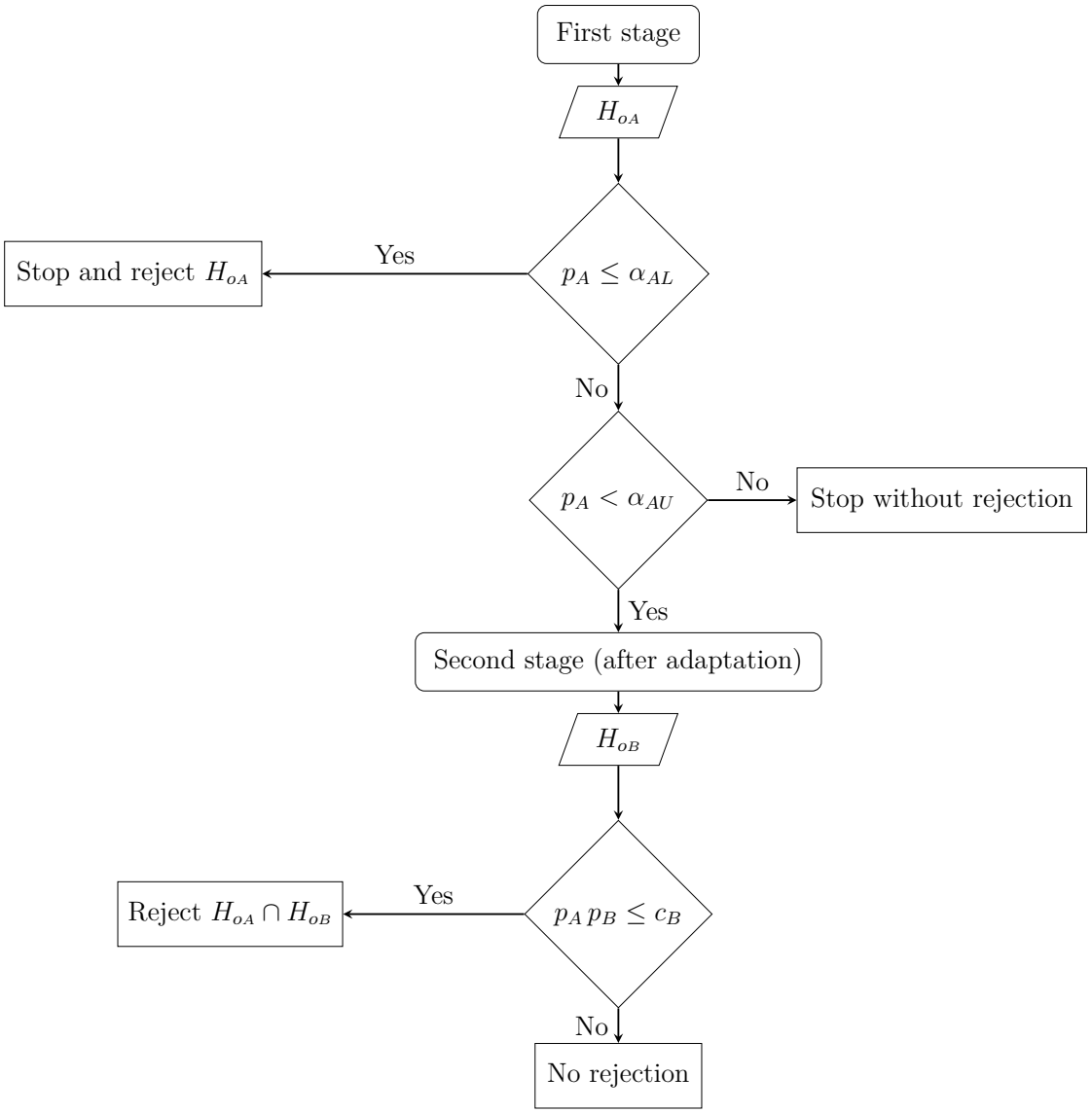


Figure 4.2: Decision making in the general seamless two-stage sequential design using Fisher's combination test approach - a flowchart visualisation. [72]

At interim analysis j:

- if there is any treatment arm with $S_{i,j} \geq u_j$, the trial is stopped and H_{oi} is rejected for all such i
- if $S_{i,j} \leq l_j$ for all treatment arms, the trial is stopped for futility
- otherwise, a subset of experimental treatments is chosen to continue along with the control to the next analysis

with $S_{i,j}$ denoting the score statistics of each treatment arm i with $i \in \{1, 2, ..., k\}$ at each interim analysis j.

A single experimental treatment to continue along with the control

Using Stallard & Todd's [61] design, only a single experimental treatment can continue to the next stage of a seamless trial—the one with the highest score statistic at the first interim analysis. Stopping boundaries $(l_1, l_2, ..., l_J)$ and $(u_1, u_2, ..., u_J)$ can be computed such that the familywise type I error is controlled in a *weak sense*:

$$P(\textit{stop and reject } H_{o(I)} \textit{ at or before interim analysis } j \mid \theta_1 = \theta_2 = ... = \theta_k = 0) = \alpha^*_U(t_j)$$

$$P(\textit{stop and do not reject } H_{o(I)} \textit{ at or before interim analysis } j \mid \theta_1 = \theta_2 = ... = \theta_k = 0) = \alpha^*_L(t_j) \tag{4.9}$$

with I denoting the experimental treatment, if any, which continues to the next stage. Note that (4.9) is equivalent to (2.4) for a seamless trial having multiple treatment arms.

As in the simple case of paragraph 2.2, an α-spending function or nominal significance levels derived using Pocock's or O'Brien & Fleming's method can be utilised to define $\alpha^*_U(t_j)$ and $\alpha^*_L(t_j)$.

To compute $(l_1, l_2, ..., l_J)$ and $(u_1, u_2, ..., u_J)$, besides (4.9), one must approximate the $(S_{I1}, S_{I2}, ..., S_{IJ})$ distribution. As in paragraph 2.2, $S_{I,j} - S_{I,j-1}$ is normally distributed with $S_{I,j} - S_{I,j-1}$ being independent of $S_{I,j-1}$. Thus, subdensities $S_{I,j}$ are given by (2.5).

Figure 2.1 (right panel) is an illustration of recursive numerical integration in a seamless design with 5 interim analyses and 4 arms (3 experimental treatments and the control), with only 2 of them (an experimental and the control) continuing to the second analysis and further testing. Note that at the first interim analysis, the distribution (and therefore the critical values) is shifted to the right to allow for treatment selection.

A pre-specified number of experimental treatments to continue along with the control

In 2008, Stallard & Friede [73] generalised Stallard & Todd's [61] method to allow for a (pre-specified) number of treatment arms to continue to the next stage of the seamless design. The method was altered so that stopping boundaries would account for the test statistics $S_{i,j}$ of the treatment arms continuing to stage II. For more information on this approach see [73].

University of Piraeus

4.4 Having more than one primary endpoint

Bauer (1991) [17] considered a design with two treatments i with $i \in \{A, B\}$ (a new treatment A versus a control B) and three primary endpoints j ($j \in \{1, 2, 3\}$) which follow a trivariate normal distribution $N(\mu^{(i)}, \Sigma)$.

Fixed sample size approach

The extended closed set of null hypotheses in the fixed sample size approach writes as follows [17]:

The null hypothesis of global efficacy:

$$H_o : \left(\mu_1^{(A)} = \mu_1^{(B)}\right) \cap \left(\mu_2^{(A)} = \mu_2^{(B)}\right) \cap \left(\mu_3^{(A)} = \mu_3^{(B)}\right) \tag{4.10}$$

The intersection null hypotheses:

$$\begin{aligned} H_{o1} &: \left(\mu_1^{(A)} = \mu_1^{(B)}\right) \cap \left(\mu_2^{(A)} = \mu_2^{(B)}\right) \\ H_{o2} &: \left(\mu_1^{(A)} = \mu_1^{(B)}\right) \cap \left(\mu_3^{(A)} = \mu_3^{(B)}\right) \\ H_{o3} &: \left(\mu_2^{(A)} = \mu_2^{(B)}\right) \cap \left(\mu_3^{(A)} = \mu_3^{(B)}\right) \end{aligned} \tag{4.11}$$

The elementary null hypotheses:

$$\begin{aligned} \widetilde{H}_{o1} &: \mu_1^{(A)} = \mu_1^{(B)} \\ \widetilde{H}_{o2} &: \mu_2^{(A)} = \mu_2^{(B)} \\ \widetilde{H}_{o3} &: \mu_3^{(A)} = \mu_3^{(B)} \end{aligned} \tag{4.12}$$

The step-down procedure dictates that:

1. H_o is tested first, at global level α.
2. If H_o is rejected, the set of intersection null hypotheses (4.11) is tested.
3. If at least two are rejected at global level α, testing proceeds to the elementary null hypotheses (4.12).

Note that one cannot test all three (or two) elementary hypotheses at global level α. If one does so, the probability of falsely rejecting at least one of them would inflate and exceed α. By rejecting the intersection hypothesis before proceeding to the elementary hypotheses, global level α is protected.

Further, if one elementary null hypothesis is rejected, then all intersection null hypotheses which involve this elementary hypothesis and the global null hypothesis are

rejected too (coherence by definition). For example, if $\widetilde{H}_{o1}$ is rejected, then H_{o1}, H_{o2} and H_o are rejected as well. On the other hand, consonance is not guaranteed (if H_o is rejected, that does not translate into rejection of any intersection or elementary hypothesis).

Testing can be performed using global test statistics. For example, H_o can be tested using a T^2-test, $\{H_{o1}, H_{o2}, H_{o3}\}$ will then be tested using T^2-tests, and subsequently elementary $\{\widetilde{H}_{o1}, \widetilde{H}_{o2}, \widetilde{H}_{o3}\}$ will be tested using t-tests, all tests at level α.

In a different line of thought, Holm (1979) [24] made good use of the union intersection principle [74] and the Bonferroni inequality, suggesting ordering the observed p-values of the elementary tests according to their magnitude and then using them as test statistics for the intersection null hypotheses. In our example with three primary endpoints:

$$p_{(1)} \leq p_{(2)} \leq p_{(3)} \tag{4.13}$$

Holm's stepwise rejective conservative procedure writes as follows [17]:

- if $(p_{(1)} \leq \alpha/3) \cap (p_{(2)} \leq \alpha/2) \cap (p_{(3)} \leq \alpha)$ then reject all elementary null hypotheses
- if $(p_{(1)} \leq \alpha/3) \cap (p_{(2)} \leq \alpha/2) \cap (p_{(3)} > \alpha)$ then reject $\widetilde{H}_{o(1)}$ and $\widetilde{H}_{o(2)}$
- if $(p_{(1)} \leq \alpha/3) \cap (p_{(2)} > \alpha/2)$ then reject $\widetilde{H}_{o(1)}$
- if $p_{(1)} > \alpha/3$ do not reject any elementary null hypothesis

The generalisation to the k elementary null hypotheses (corresponding to k primary endpoints) is straightforward:

- if $(p_{(1)} \leq \alpha/k) \cap (p_{(2)} \leq \alpha/(k-1)) \cap ... \cap (p_{(k)} \leq \alpha)$ then reject all elementary null hypotheses
- if $(p_{(1)} \leq \alpha/k) \cap (p_{(2)} \leq \alpha/(k-1)) \cap ... \cap (p_{(k)} > \alpha)$ then reject all intersection null hypotheses but those involving the elementary null hypothesis of $\widetilde{H}_{ok}$ and so on
- last, if $p_{(1)} > \alpha/k$ do not reject any elementary null hypothesis

Further, elementary null hypotheses can be weighed using different multiplicators for the observed p-values. [24][75]

Instead of using the Bonferroni inequality, Rüger (1978) [76] proposed a less conservative approach. He suggested H_o to be rejected if $p_{(r)} \leq r\alpha/k$ with r being fixed in advance $1 < r \leq k$. For further information on approaches for the fixed sample size design, the interested reader may refer to work by Hommel (1986, 1988) [75][77], Simes (1986) [27] and Abt (1987) [78].

University of Piraeus

Sequential design

Bauer (1986) [79] investigated a two-stage design with two treatment arms i and two primary endpoints $j-$one efficacy endpoint (stopping for efficacy) and the toxicity rate (stopping for toxicity). He assumed that the two variables $(Z_1^{(i)}, Z_2^{(i)})$ follow a bivariate, normal distribution $N(\mu^{(i)}, \Sigma)$ and have an unknown correlation ρ_i. The test statistics for each stage and treatment arm are denoted as:

$$Y^T = (Y_{1n}, Y_{2n}, Y_{1N}, Y_{2N}) \tag{4.14}$$

with n the sample size at the end of the first stage, and N the total sample size. The test statistics are given by:

$$Y_{1M} = \frac{1}{\sqrt{2M}} \left(\sum_{m=1}^{M} Z_{lm}^{(1)} - \sum_{m=1}^{M} Z_{lm}^{(2)} \right) \tag{4.15}$$

with $l \in \{1, 2\}$ and $M \in \{n, N\}$. Note that $Z_{lm}^{(i)}$ denotes the co-primary variable response of each treatment arm i.

Y follows a multivariate normal distribution with a mean of:

$$\mu_Y^T = \left(\delta_1 \sqrt{(\frac{n}{2})},\ \delta_2 \sqrt{(\frac{n}{2})},\ \delta_1 \sqrt{(\frac{N}{2})},\ \delta_2 \sqrt{(\frac{N}{2})} \right) \tag{4.16}$$

where $\delta_1 = \mu_1^{(1)} - \mu_1^{(2)}$ and $\delta_2 = \mu_2^{(1)} - \mu_2^{(2)}$, and a covariance matrix of:

$$\Sigma = \begin{pmatrix} 1 & \frac{\rho_1+\rho_2}{2} & \sqrt{\frac{n}{N}} & \sqrt{\frac{n}{N}}\frac{\rho_1+\rho_2}{2} \\ & 1 & \sqrt{\frac{n}{N}}\frac{\rho_1+\rho_2}{2} & \sqrt{\frac{n}{N}} \\ & & 1 & \frac{\rho_1+\rho_2}{2} \\ & & & 1 \end{pmatrix} \tag{4.17}$$

The null hypothesis of interest writes as follows:

$$H_o : \delta^T = (\delta_1, \delta_2)^T = (0, 0)^T \equiv H_{o1} \cap H_{o2} \tag{4.18}$$

with

$$\begin{aligned} H_{o1} &: \delta_1 = 0 \\ H_{o2} &: \delta_2 = 0 \end{aligned} \tag{4.19}$$

H_o is rejected if at least one individual hypothesis (H_{o1}, H_{o2}) is rejected. Critical regions W_k for the individual test statistics $\{Y_{1n}, Y_{2n}, Y_{1N}, Y_{2N}\} = \{Y_1, Y_2, Y_3, Y_4\}$ are chosen so that:

$$P\left(\bigcap_{k=1}^{4}(Y_k \notin W_k)|H_o\right) = 1-\alpha \tag{4.20}$$

For the one-sided test:

$$W_k \equiv (c_k, +\infty) \tag{4.21}$$

For the symmetric two-sided test:

$$W_k \equiv (-\infty, -c_k) \cap (c_k, +\infty) \tag{4.22}$$

The problem now can be summarised as:

How to choose W_k(or c_k), and therefore the "individual" marginal significance levels $\alpha_k = P(Y_k \in W_k|H_o)$, in different test situations when the only known condition is (4.20)?

If $\alpha_2 = \alpha_4 = 0$ ($\alpha_1 = \alpha_3 = 0$), the problem would be identical to a repeated significance testing for the efficacy endpoint (toxicity endpoint) only. For $\alpha_3 = \alpha_4 = 0$ ($\alpha_1 = \alpha_2 = 0$), the problem would reduce to a fixed sample test of sample size $n\,(N)$ and two primary endpoints. If 3 out of 4 α_k were zero, the problem would reduce to a fixed sample size test for a single random variable. If $0 < \alpha_k < 1$ for all (or some) $k \in \{1, 2, 3, 4\}$ then various scenarios could arise.

Scenario 1:

- *The procedure stops at the first stage due to unacceptable toxicity if* $(Y_1 \notin W_1) \cap (Y_2 \in W_2)$.

Scenario 2:

- *The procedure stops at the first stage for efficacy reason if* $(Y_1 \in W_1) \cap (Y_2 \notin W_2)$.

Scenario 3:

- *The procedure stops at the first stage due to unacceptable toxicity if* $(Y_1 \in W_1) \cap (Y_2 \in W_2)$.

Scenario 4:

- *The procedure continues to the second stage if* $(Y_1 \notin W_1) \cap (Y_2 \notin W_2)$

To determine the stopping boundaries c_k, the problem of endpoint correlation must be tackled. Bauer [79] adopted results known from the multivariate normal distribution. For the two-sided symmetrical case [80]:

$$P\left((Y_{(1)} \in C_1) \cap (Y_{(2)} \in C_2)\right) \geq P(Y_{(1)} \in C_1)\, P(Y_{(2)} \in C_2) \tag{4.23}$$

with $Y_{(1)}^T = (Y_1, Y_2, ..., Y_r)$ and $Y_{(2)}^T = (Y_{r+1}, Y_{r+2}, ..., Y_s)$.

From (4.20) and (4.23), if

$$P\left((Y_{(1)} \notin W_1) \cap (Y_{(3)} \notin W_3)\,|H_o\right)\; P\left((Y_{(2)} \notin W_2) \cap (Y_{(4)} \notin W_4)\,|H_o\right) = 1-\alpha \tag{4.24}$$

it stands that

$$P\left(\bigcap_{k=1}^{4}(Y_k \notin W_k)|H_o\right) \not< 1-\alpha \tag{4.25}$$

Following this line of thought, one can use (4.24) instead of (4.20). For the one-sided test, Bauer [79] using results by Sleplan (1962) [81] showed that if $\rho_1+\rho_2>0$ (a condition usually fulfilled), (4.24) can again be applied instead of (4.20), and the equality sign in (4.20) can now be replaced by $\geq$.

Unfortunately (4.24) alone does not suffice to determine the stopping boundaries c_k. A way to approach the problem is as follows [79]:

For the one-sided test, under the pure location shift alternative:

- $(\delta_1,\delta_2)^T \equiv (\delta_{11},\delta_{21})^T$ with $0 \leq \delta_{21} \leq \delta_{11}$ being critical values for δ_2 and δ_1 respectively, the probability of stopping at the first or the second stage with H_{o1} rejection is γ_1, with γ_1 being equal to:

$$\gamma_1 = P_1 = P\left((Y_1 \in W_1) \cup [(Y_1 \notin W_1) \cap (Y_2 \notin W_2) \cap (Y_3 \in W_3)]\,|\,\delta_{11},\delta_{21}\right) \tag{4.26}$$

- $(\delta_1,\delta_2)^T \equiv (\delta_{12},\delta_{22})^T$ with $0 \leq \delta_{12} \leq \delta_{22}$ being critical values for δ_1 and δ_2 respectively, the probability of stopping at the first or the second stage with H_{o2} rejection is γ_2, with γ_2 being equal to:

$$\gamma_2 = P_2 = P\left((Y_2 \in W_2) \cup [(Y_1 \notin W_1) \cap (Y_2 \notin W_2) \cap (Y_4 \in W_4)]\,|\,\delta_{12},\delta_{22}\right) \tag{4.27}$$

For the one-sided test:

- if $\delta_1 > \delta_{11}$ and $\delta_2 < \delta_{21}$ then $P_1 > \gamma_1$
- if $\delta_1 < \delta_{12}$ and $\delta_2 > \delta_{22}$ then $P_2 > \gamma_2$

Note that (4.26) and (4.27) depend on the unknown ρ, hence, the impact of the correlation between the co-primary endpoints has still to be determined.

Equations (4.26) and (4.27) have to be modified in case of a two-sided test. More specifically, $Y_i \in W_i$ has to be replaced by $Y_i \in (c_i,+\infty)$ because stopping would occur when the test statistics fall into $(c_k,+\infty)$ and not into $(-\infty,-c_k)$.

Again (4.24) with the extra conditions (4.26) and (4.27) does not suffice to determine the critical values c_k, and subsequently the nominal levels α_k. Further

assumptions shall be made with regard to sample size n, means difference δ, power γ and nominal levels α_k. Bauer [79] investigated a case of equal sample sizes for each stage: $n = N/2$, the means difference of the two primary endpoints has the same magnitude: $\delta_{11} = \delta_1^* = \delta$, $\delta_{22} = \delta_2^* = \delta$, $\delta_{21} = \delta_{12} = 0$, and under the alternative, power $\gamma = \gamma_1 = \gamma_2$. He also added a fourth condition to (4.24), (4.26) and (4.27): same nominal levels for the two primary endpoints at the first stage $\alpha_1 = \alpha_2$ (symmetry implies that for the second stage $\alpha_3 = \alpha_4$). Results are shown in Table 4.1 (for $\delta_1^*/\delta_2^* = 1$). For $\gamma = 0.90$, $\alpha_1 = \alpha_2 = 0.015$; this result does not change even when correlation is as large as $\rho = 0.90$ (see also Table 4.2).

δ_1^*	δ_2^*	δ_1^*/δ_2^*		$\gamma = 0.80$		$\gamma = 0.90$	
				(1)	(2)	(1)	(2)
0.25	0.25	1	N	282	340	376	440
			α_1	0.015	0.015	0.015	0.015
			α_2	0.015	0.015	0.015	0.015
0.25	0.33	0.76	N	232	286	314	376
			α_1	0.027	0.027	0.028	0.028
			α_2	0.0030	0.0018	0.0021	0.0012
			(c_2)	(2.75)	(3.12)	(2.86)	(3.23)
0.33	0.5	0.66	N	128	158	174	210
			α_1	0.029	0.029	0.030	0.029
			α_2	<0.001	<0.001	<0.001	<0.001
			(c_2)	(3.13)	(3.57)	(3.31)	(3.75)
0.25	0.5	0.50	N	222	280	306	370
			α_1	0.030	0.029	0.030	0.029
			α_2	<0.001	<0.001	<0.001	<0.001
			(c_2)	(4.35)	(5.0)	(4.8)	(5.4)

Table 4.1: A two-stage group sequential design having two treatment arms and two co-primary endpoints (an efficacy and a toxicity endpoint)−required total sample size (N), nominal significance level for the efficacy endpoint (α_1) and the toxicity endpoint (α_2), and stopping boundary for the toxicity endpoint (c_2), for two different values of power (γ) and various combinations of means difference (δ_1^*, δ_2^*) assuming zero correlation between endpoints ($\rho = 0$) (global $\alpha = 0.05$). (1)=one-sided test, (2)=two-sided test. [79]

ity of Piraeus University

γ	$\rho = 0$	$\rho = 0.9$
0.80	340	336
0.90	440	436
0.95	536	532

Table 4.2: Sensitivity analysis of the trial design presented in Table 5.1 (with $\delta_1^*/\delta_2^* = 1$): required total sample size for different values of power (γ) and two extreme values of endpoint correlation ρ. [79]

Maximum sample size can be estimated by:

$$N_{max} = \left(\frac{\delta_1^*}{\delta_2^*}\right)^2 N \tag{4.28}$$

Comparison of various designs

5.1 Introduction

In this chapter, we attempt to compare various designs with regard to the total sample size needed to reach the expected statistical power, and their ability to detect a treatment difference, if one exists, engaging fewer patients. The best designs are chosen based on the expected sample size, the cumulative number of patients at the interims, the total trial duration, and the possibility of stopping for efficacy if a treatment difference actually exists.

Of note, all the trial designs and the treatment schemes presented here are not real.

5.2 Methodology

The parameters that change among the models are: the spending function parameters, the number and timing of the interim looks.

Group sequential design with a time-to-event endpoint

The input parameters for a design utilising a time-to-event endpoint to terminate the trial are the hazards ratio under the alternative, the expected median survival time of the control arm, the type I error & the statistical power, the expected censoring rate, the expected enrolment ramp-up duration, the expected accrual rate after ramp-up phase, the minimum follow-up duration for all patients before the final analysis, the number of interim looks, the interval spacing, and the spending function parameters. The output design parameters are: the total sample size, the cumulative sample size at each interim, the cumulative number of events, the total trial duration, the stopping boundaries, and the cumulative crossing probabilities.

University of Piraeus

Group sequential design with a binary endpoint

The input parameters for a design utilising a binary endpoint are: the event rate of each arm under the alternative, the type I error & the statistical power, the sample size of the fixed design, the number of interim looks, the interval spacing, and the spending function parameters. The output design parameters are the following: the total sample size, the cumulative sample size at each interim, the stopping boundaries, and the cumulative crossing probabilities.

Software

The design simulation was performed in *RStudio* (*gsDesign R* package). Additionally, *gsDesign Explorer v0.61* was explored. *gsDesign explorer* is a web-based interface to the open-source *gsDesign R* package for designing group sequential trials built using the *RStudio Shiny* package. *Creator & Project Manager: Keaven Anderson; Shiny Developer: John Lueders. Url:* https://gsdesign.shinyapps.io/prod/

5.3 ELPIDA: A two-arm design utilising a time-to-event endpoint

ELPIDA is a multicentre, double blind, phase III trial exploring the addition of a new monoclonal antibody to the standard of care for women with advanced, previously untreated triple negative breast cancer (TNBC). Eligible patients will be randomised 1:1 to receive either stereostatic body radiation (SBRT) to all lesion sites followed by 4-5 cycles (Q3W) of standard of care chemotherapy (SOCC) (arm A) or SBRT followed by 4-5 SOCC cycles (Q3W) and adjuvant *Elpizumab* (arm B). After the end of SOCC, all patients without progression will undergo surgery. Arm B will receive *Elpizumab* (Q4W) 3 weeks ($\pm$ 1 week) after surgery and will continue for one year or until first disease progression, intolerance or refusal, whichever comes first. After surgery, arm A will receive placebo in an analogous treatment scheme. Patients with CNS metastasis, active autoimmune disease, pregnant or breastfeeding women are not eligible in this trial.

Elpizumab is considered a quite promising treatment as it is expected to lower risk of death to 65% as compared to the standard of care (HR: 0.65). From historical data, median overall survival (OS) of women with advanced, previously untreated TNBC receiving the standard of care is 15.1 months.

Statistical concerns

Patients are gradually enrolled in different sites, and therefore a group sequential scheme is utilised. In this superiority trial design, type I error & statistical power are set equal to 0.025 (one-sided) and 80%, respectively.

The Kim & Tsiatis's (1990) [82] method will be used to determine the trial duration. Using this method, the enrolment rate and the follow-up duration will be fixed, and the total trial duration will be determined to power the design. After a ramp-up period of 6 months, 22 patients per month are expected to be recruited at the 8 centres. Final analysis is expected to be performed 6 months after randomisation of the last patient.

Finally, an exponential drop-out (censoring) rate of 5% is expected in the trial (indicating the rate of lost to follow-up patients).

University of Piraeus

Scenario A

Two equally spaced looks are planned. Efficacy and futility bounds are derived using a Lan & DeMets's spending function approximating O'Brien & Fleming's bound.

Scenario B

Three looks are planned, at 30%, 60% and 100% of the information time. Efficacy and futility bounds are derived using a Lan & DeMets's spending function approximating O'Brien & Fleming's bound.

Scenario C

Three looks are planned, at 30%, 60% and 100% of the information time. Efficacy and futility bounds are derived using a Lan & DeMets's spending function approximating O'Brien & Fleming's and Pocock's bound, respectively.

Scenario	t_1	t_2	t_3	Lower bound	Upper bound
A	0.50	1.0		O'Brien&Fleming	O'Brien&Fleming
B	0.30	0.60	1.0	O'Brien&Fleming	O'Brien&Fleming
C	0.30	0.60	1.0	Pocock	O'Brien&Fleming

Table 5.1: Design parameters of the 3 scenarios. t_1, t_2 and t_3 denote the time points that the interim looks are planned.

Fixed design

The sample size of a trial having only one analysis would be 527 patients. A total of 170 events would be required to reach the expected power of 80%. The trial duration would be 33.0 months, whilst the enrolment is expected to last for 27.0 months.

Results - Scenario A

A total of 550 patients have to be recruited to detect a hazards ratio of 0.65. In total, 179 events are needed to reach the required 80% power. First analysis will be performed after half of them (90 events) have been reached, engaging 410 patients (74.5%). Enrolment and total study duration are expected to be 28 and 34 months, respectively.

From Figure 5.1, it is clear that:

- at interim 1, the trial will be terminated with H_o rejection if the normal test statistic exceeds the value of **2.96**
- at interim 1, the trial will be terminated for futility if the normal test statistic is less than **0.56**
- at the final analysis, H_o will be rejected if the normal test statistic exceeds the value of **1.97**

This is also graphically depicted in Figure 5.2. Also:

- $P(\text{stopping for efficacy at interim } 1 \,|\, H_o) = 0.0015$
- $P(\text{stopping for futility at interim } 1 \,|\, H_o) = 0.71$
- $P(\text{stopping for efficacy at interim } 1 \,|\, H_1) = 0.18$
- $P(\text{stopping for futility at interim } 1 \,|\, H_1) = 0.070$

- $P(\text{stopping for efficacy at or before the final analysis } |\, H_o) = 0.023 \quad (= \alpha)$
- $P(\text{stopping for futility at or before the final analysis } |\, H_o) = 0.98 \quad (= 1 - \alpha)$
- $P(\text{stopping for efficacy at or before the final analysis } |\, H_1) = 0.80 \quad (= 1 - \beta)$
- $P(\text{stopping for futility at or before the final analysis } |\, H_1) = 0.20 \quad (= \beta)$

Additionally, the model predicts that if the observed HR at the first look is ≤ 0.53 the boundary will be crossed and the trial will be stopped for efficacy. On the other hand, if the observed HR at the first look is ≥ 0.89 the trial will be stopped for futility.

Figure 5.1 also shows the expected time in months that the interim analyses are expected to occur. More precisely, the model predicts that the 90 events needed for the first analysis will be reached at 21.6 months from randomisation of the first patient. However, this is highly dependent on the observed event rate, and thus reality might be quite different as the trial goes on. In case of high event rate, the 90 events will occur earlier than expected and therefore the data will be inspected sooner. In case of extremely low event rate, the review board might even decide to close the trial.

University of Piraeus

Analysis	Value	Efficacy	Futility
IA 1: 50%	Z	2.9626	0.5594
N: 410	p (1-sided)	0.0015	0.2879
Events: 90	HR at bound	0.5345	0.8884
Month: 21.6	P(Cross) if HR=1	0.0015	0.7121
	P(Cross) if HR=0.65	0.1770	0.0699
Final	Z	1.9686	1.9686
N: 550	p (1-sided)	0.0245	0.0245
Events: 179	HR at bound	0.7450	0.7450
Month: 34	P(Cross) if HR=1	0.0233	0.9767
	P(Cross) if HR=0.65	0.8000	0.2000

Figure 5.1: Tabular summary of scenario A. An asymmetric, two-sided scheme was utilised for the boundaries. Efficacy and futility bounds were set using Lan-DeMets spending function approximating O'Brien & Fleming's bound. Hazards ratio (HR) presented here is not a requirement, but an estimate of the HR required to cross each bound. Month is estimated given enrolment and event rate assumptions under the alternate hypothesis. P(Cross) is the probability of crossing the given bound (efficacy or futility) at or before the given analysis under the assumed HR. Design assumes futility bound is discretionary (non-binding). Image taken using the gsDesign Explorer v0.61.

In Figure 5.3, the black solid line represents the cumulative crossing probability at each interim vs. true HR. Notice that above the value of 0.60, the probability of rejecting H_o rapidly decreases. As expected, the probability of rejecting at the first look under the alternative (dashed black line) is much lower.

Figure 5.4 shows how type I and type II errors are spent with information time (expected events at time t / total expected events).

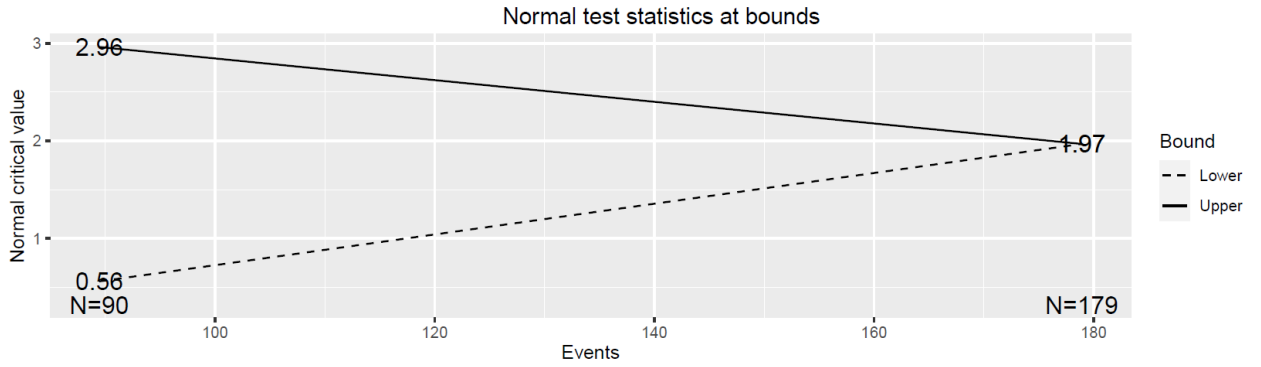


Figure 5.2: Stopping boundaries for scenario A. The solid (dashed) line represents the efficacy (futility) boundary.

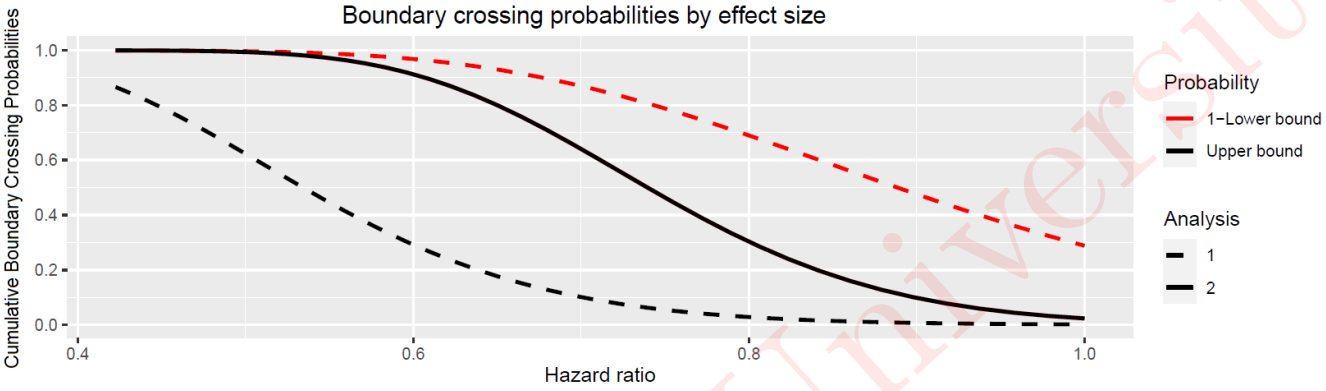


Figure 5.3: Cumulative boundary crossing probabilities by effect size for scenario A. Power by effect size is represented by the solid black line. Power at the first look is represented by the black dashed line. One minus the probability of crossing the futility bound by the first look is represented by the red dashed line.

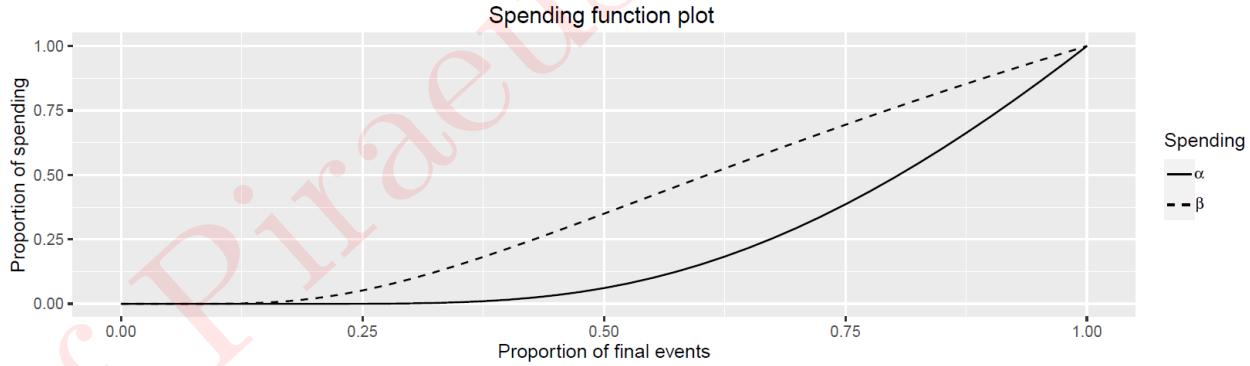


Figure 5.4: Alpha & beta -spending functions (solid and dashed line, respectively) approximating O'Brien & Fleming's boundary.

University of Piraeus

Results - Scenario B

In total, 564 patients have to be recruited to detect a hazards ratio of 0.65. 184 events are needed to reach the required 80% power. First analysis will be performed after 56 events have been reached, engaging 304 patients (53.9%). Second analysis will be performed after 111 events, engaging 470 patients (83.3%). Enrolment and total study duration are expected to be 28.6 and 34.6 months, respectively.

Analysis	Value	Efficacy	Futility
IA 1: 30%	Z	3.9286	-0.4699
N: 304	p (1-sided)	0.0000	0.6808
Events: 56	HR at bound	0.3472	1.1349
Month: 16.8	P(Cross) if HR=1	0.0000	0.3192
	P(Cross) if HR=0.65	0.0099	0.0193
IA 2: 60%	Z	2.6700	0.9338
N: 470	p (1-sided)	0.0038	0.1752
Events: 111	HR at bound	0.6015	0.8371
Month: 24.3	P(Cross) if HR=1	0.0038	0.8289
	P(Cross) if HR=0.65	0.3414	0.0980
Final	Z	1.9810	1.9810
N: 564	p (1-sided)	0.0238	0.0238
Events: 184	HR at bound	0.7466	0.7466
Month: 34.6	P(Cross) if HR=1	0.0225	0.9775
	P(Cross) if HR=0.65	0.8000	0.2000

Figure 5.5: Tabular summary of scenario B. An asymmetric, two-sided scheme was utilised for the boundaries. Efficacy and futility bounds were set using a Lan & DeMets's spending function approximating O'Brien & Fleming's bound. Hazards ratio (HR) presented here is not a requirement, but an estimate of the HR required to cross each bound. Month is estimated given enrolment and event rate assumptions under the alternate hypothesis. P(Cross) is the probability of crossing the given bound (efficacy or futility) at or before the given analysis under the assumed HR. Design assumes futility bound is discretionary (non-binding). Image taken using the gsDesign Explorer v0.61.

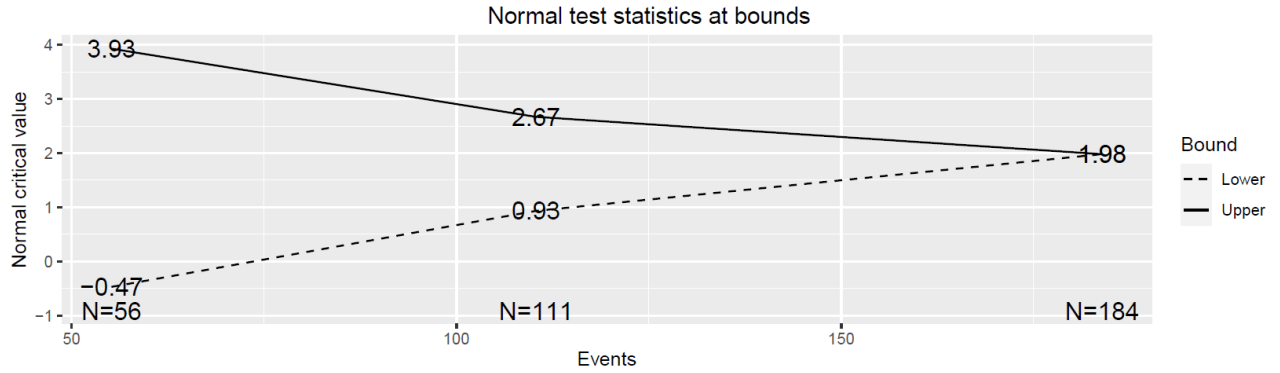


Figure 5.6: Stopping boundaries for scenario B. The solid (dashed) line represents the efficacy (futility) boundary.

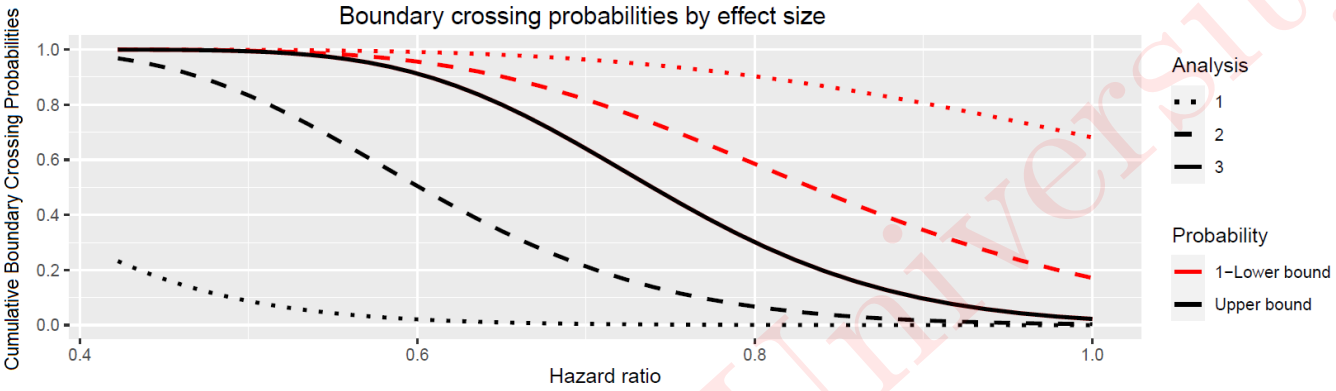


Figure 5.7: Cumulative boundary crossing probabilities by effect size for scenario B. Power by effect size is represented by the solid black line. Power at interim 1 is represented by the black dotted line. One minus the probability of crossing the futility bound by interim 1 is represented by the red dotted line. Power at interim 2 is represented by the black dashed line. One minus the probability of crossing the futility bound by interim 2 is represented by the red dashed line.

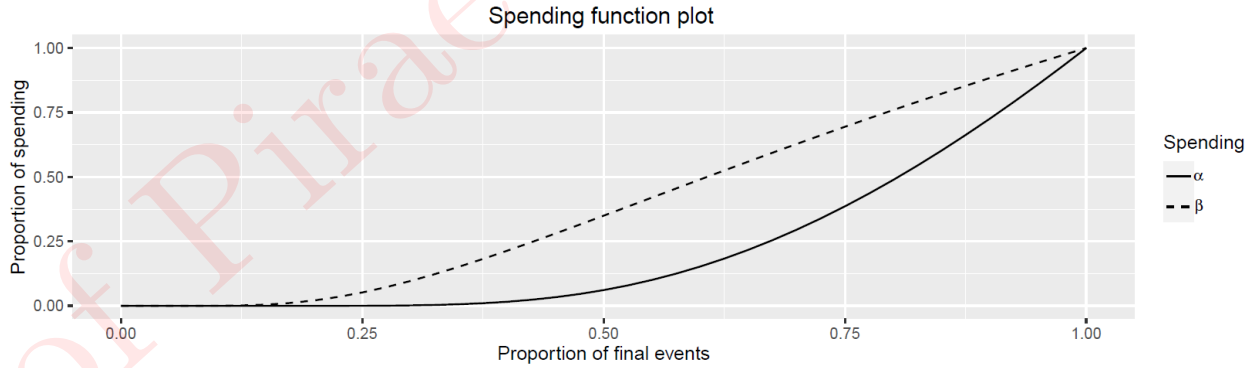


Figure 5.8: Alpha & beta -spending functions (solid and dashed line, respectively) approximating O'Brien & Fleming's boundary.

University of Piraeus

Results - Scenario C

In total, 622 patients have to be recruited to detect a hazards ratio of 0.65. 208 events are needed to reach the required 80% power. First analysis will be performed after 63 events have been reached, engaging 328 patients (52.7%). Second analysis will be performed after 125 events, engaging 508 patients (81.7%). Enrolment and total study duration are expected to be 31.2 and 37.2 months, respectively.

Analysis	Value	Efficacy	Futility
IA 1: 30%	Z	3.9286	0.3134
N: 328	p (1-sided)	0.0000	0.3770
Events: 63	HR at bound	0.3693	0.9236
Month: 17.8	P(Cross) if HR=1	0.0000	0.6230
	P(Cross) if HR=0.65	0.0128	0.0831
IA 2: 60%	Z	2.6700	1.1240
N: 508	p (1-sided)	0.0038	0.1305
Events: 125	HR at bound	0.6196	0.8175
Month: 26.1	P(Cross) if HR=1	0.0038	0.8874
	P(Cross) if HR=0.65	0.3926	0.1417
Final	Z	1.9810	1.9810
N: 622	p (1-sided)	0.0238	0.0238
Events: 208	HR at bound	0.7595	0.7595
Month: 37.2	P(Cross) if HR=1	0.0197	0.9803
	P(Cross) if HR=0.65	0.8000	0.2000

Figure 5.9: Tabular summary of scenario C. An asymmetric, two-sided scheme was utilised for the boundaries. Efficacy and futility bounds were set using a Lan & DeMets's spending function approximating O'Brien & Fleming's and Pocock's bound, respectively. Hazards ratio (HR) presented here is not a requirement, but an estimate of the HR required to cross each bound. Month is estimated given enrolment and event rate assumptions under the alternate hypothesis. P(Cross) is the probability of crossing the given bound (efficacy or futility) at or before the given analysis under the assumed HR. Design assumes futility bound is discretionary (non-binding). Image taken using the gsDesign Explorer v0.61.

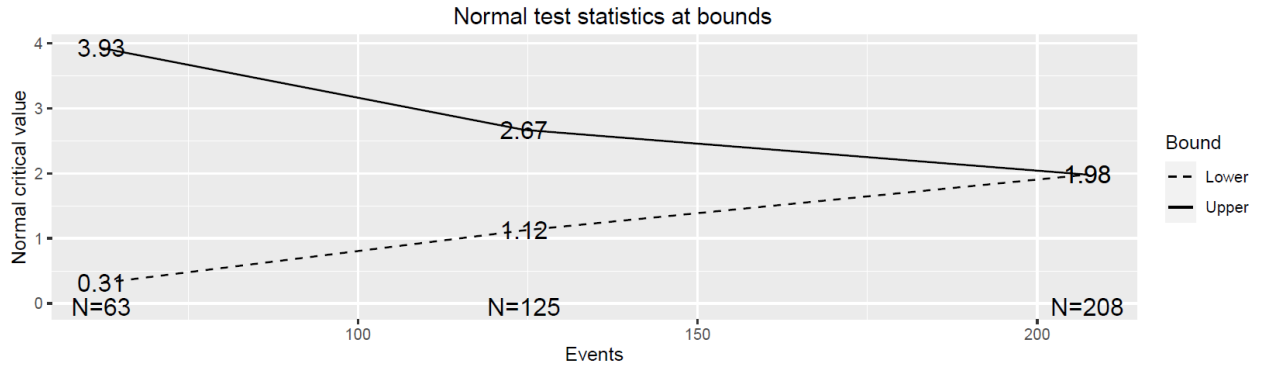


Figure 5.10: Stopping boundaries for scenario C. The solid (dashed) line represents the efficacy (futility) boundary.

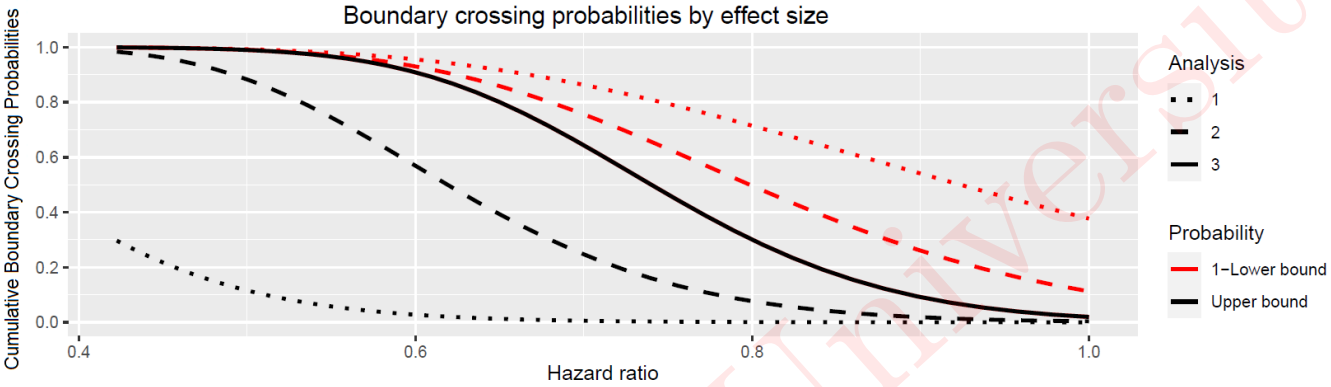


Figure 5.11: Cumulative boundary crossing probabilities by effect size for scenario C. Power by effect size is represented by the solid black line. Power at interim 1 is represented by the black dotted line. One minus the probability of crossing the futility bound by interim 1 is represented by the red dotted line. Power at interim 2 is represented by the black dashed line. One minus the probability of crossing the futility bound by interim 2 is represented by the red dashed line.

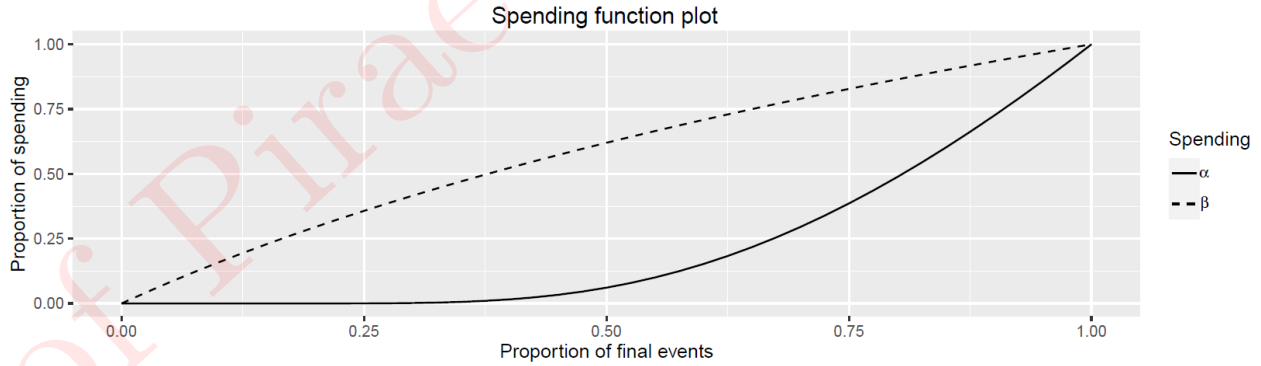


Figure 5.12: Alpha-spending function (solid line) approximating O'Brien & Fleming's bound. Beta-spending function (dashed line) approximating Pocock's bound.

University of Piraeus

5.4 Discussion on ELPIDA trial

As expected, sample size of all the sequential designs is greater than the fixed one (with 527 required patients) as the sequential design slightly increases the total sample size under H_o. Although the 3 designs demand 23 (a 4.4% increase), 37 (7.0%) and 95 (18%) more patients than the fixed one, they reap all the benefits of a group sequential design, and thus they are much preferred to a fixed design. From all 3 scenarios, the one with the smallest sample size is A with 550 patients. Attained α is 0.023 for A & B, and 0.020 for C.

Although design B requires 14 more patients than A, it has an advantage upon A with regard to its ability to detect a treatment difference, if one exists. More precisely, under the alternative, H_o will be rejected at or before the second look with 34% chance. If H_o is rejected at the second interim, 470 patients of the total 564 required (83.3%) would be needed to terminate the trial in favour of the experimental treatment (10.8% less patients than the fixed design). Design A, on the other hand, will reject H_o, if H_o is not true, with 18% chance at the interim look, engaging 410 patients of the total 550 required (74.5%). Additionally, the chance of stopping for futility under H_o amounts to 71% at the interim, while the respective value for B is 32% at the first look (engaging only 304 patients) and 83% at the second. The trial duration needed in B is longer only by 0.6 month as compared to A (and only 1.6 month longer than the fixed design).

Despite the aforementioned, design A has a great drawback compared to other designs. The model predicts that the 90 events required for the first analysis (50% of the information time) will be reached only after 410 patients have been enrolled. It is clear that 410 patients is a large number of patients to be involved before looking at the data.

Design C, on the other hand, demands the largest number of patients. Its biggest advantage upon B is the ability to terminate the trial for futility under H_o with 62% chance at the first look engaging only 328 patients. However, its large sample size (95 more patients that the fixed design) is still a deterrent factor.

Following this line of thought, the biometrician searches for other scenarios that would probably yield better results. Figure 5.13(I) depicts the crossing probabilities and the sample size at different time points (t_i) of the interim look for a scenario having 2 looks and utilising the O'Brien & Fleming's bound. Notice that for t=0.50, we get design A. Designs with t $\geq$0.50 would not be chosen due to the fact that they require more than 400 patients at the interim, while designs with t $\leq$0.40

are not suitable due to their decreased ability to detect a treatment difference at the interim, if one exists (<6% chance). The best-case scenario for an O'Brien & Fleming's design with two looks is to plan the interim at 45% of the information time (scenario D; results presented in Table 5.3). Design D engages 546 patients (4 less than A), and demands 382 patients at the interim, at the expense of a lower chance of stopping for futility under H_o (63%) (but still satisfactory), and a lower chance of stopping for efficacy under the alternative (11%). Attained α is 0.024.

A analogous simulation has been performed for a scenario with 3 looks (Figure 5.13 IIa-c). This situation is more complex as two parameters have to be optimised (t_1 and t_2 for the first and second look, respectively). If we demand a sample size of less than 340 patients at the interim and >40% chance of stopping for futility under H_o at the first look, the scenario which derives the smallest total sample size is the one with (t_1=0.35, t_2=0.60; scenario E). Design E demands the same sample size as B and engages 334 patients at the interim (30 more than B), and has an attained α of 0.023. However, it offers a greater chance of stopping for futility under H_o (44%). As expected, the crossing probabilities at or before the second look remain the same (the timing of the second look has not been changed).

Conclusion on ELPIDA trial

The best-case scenarios with regard to the total expected sample size and the crossing probabilities are scenarios D&E. Between the two, the safest choice would be E, as the large sample size that ELPIDA trial demands (527 for the fixed design) suggests to plan for more than one look before the final analysis. However, the final choice will be taken in collaboration with the principal investigator and the multidisciplinary team of the trial. Both designs use an α-spending function approximating O'Brien & Fleming's bound. Pocock's bound (scenario C) was rejected due to the increased sample size that this approach derives. The O' Brien & Fleming's bound is widely accepted in clinical research.

ity of Piraeus University

Scenario	t_1	t_2	t_3	Lower bound	Upper bound
A	0.50	1.0		O'Brien&Fleming	O'Brien&Fleming
B	0.30	0.60	1.0	O'Brien&Fleming	O'Brien&Fleming
C	0.30	0.60	1.0	Pocock	O'Brien&Fleming
D	0.45	1.0		O'Brien&Fleming	O'Brien&Fleming
E	0.35	0.60	1.0	O'Brien&Fleming	O'Brien&Fleming

Table 5.2: Design parameters of all 5 scenarios of ELPIDA trial. Highlighted entries represent new scenarios.

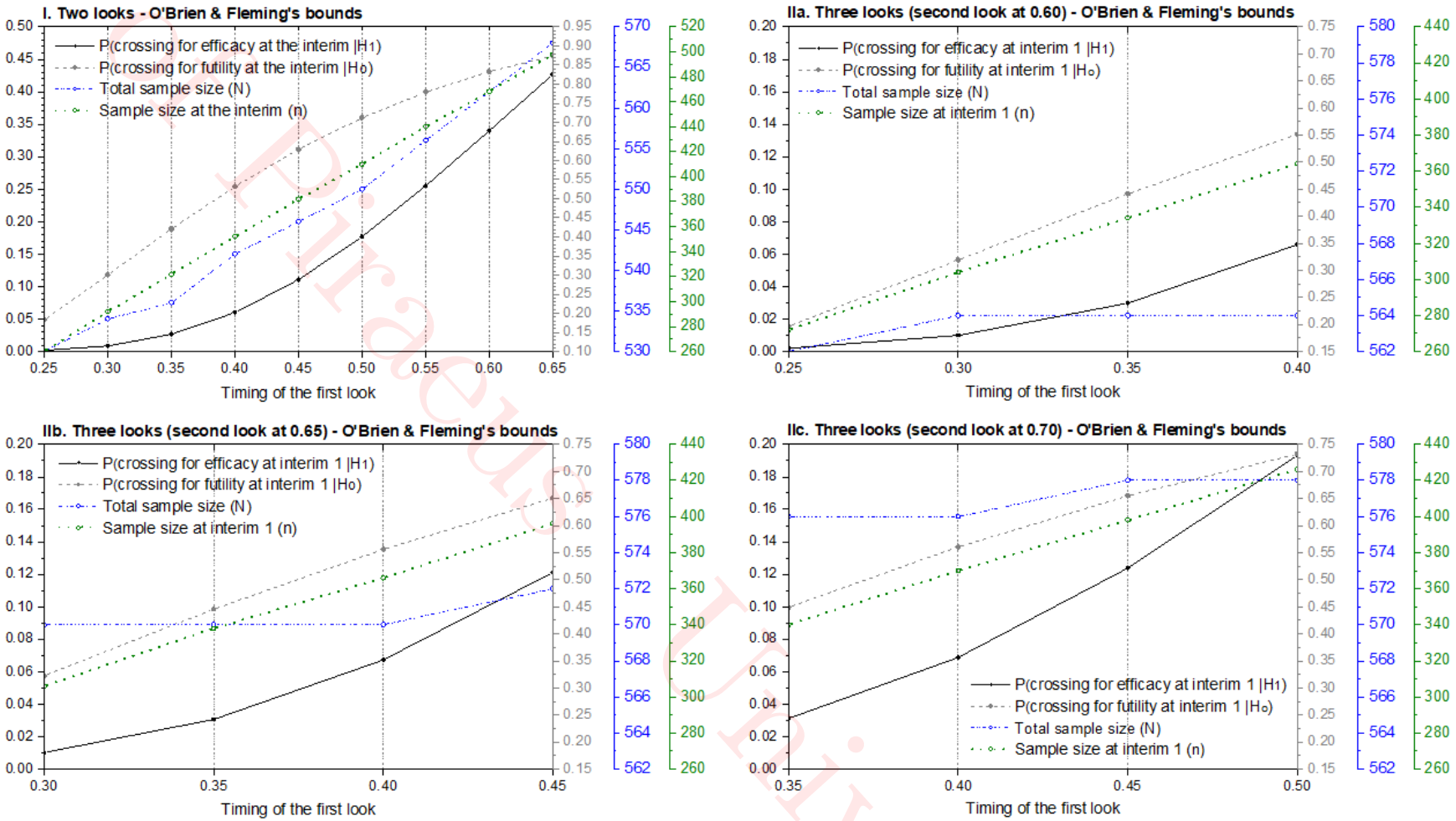


Figure 5.13: Crossing probabilities and sample size for several scenarios of ELPIDA trial. I) A design with two looks. IIa-c) A design with three looks—second look at 60% (IIa), 65% (IIb) and 70% (IIc) of the information time. All designs utilise the O'Brien & Fleming's bound. Results are presented as a function of the timing of the first look.

University of Piraeus Library

	Scenario					
	A	B	C	D	E	Fixed
N(Events) - Overall	550 (179)	564 (184)	622 (208)	546 (177)	564 (185)	527 (170)
n(events) at interim 1	410 (90)	304 (56)	328 (63)	382 (80)	334 (65)	*NA*
n(events) at interim 2	*NA*	470 (111)	508 (125)	*NA*	470 (111)	*NA*
Total trial duration (in months)	34.0	34.6	37.2	33.8	34.6	33.0
P(stop for efficacy at interim 1| H_o)	0.0015	<0.001	<0.001	<0.001	<0.001	*NA*
P(stop for futility at interim 1| H_o)	0.71	0.32	0.62	0.63	0.44	*NA*
P(stop for efficacy at interim 1| H_1)	0.18	0.010	0.013	0.11	0.030	*NA*
P(stop for futility at interim 1| H_1)	0.070	0.019	0.083	0.056	0.030	*NA*
P(stop for efficacy at or before interim 2| H_o)	*NA*	0.0038	0.0038	*NA*	0.0038	*NA*
P(stop for futility at or before interim 2| H_o)	*NA*	0.83	0.89	*NA*	0.83	*NA*
P(stop for efficacy at or before interim 2| H_1)	*NA*	0.34	0.39	*NA*	0.34	*NA*
P(stop for futility at or before interim 2| H_1)	*NA*	0.098	0.14	*NA*	0.098	*NA*

Table 5.3: Output design parameters and cumulative crossing probabilities of ELPIDA trial (all scenarios).

5.5 BREATH: A two-arm design utilising a binary endpoint

BREATH is a multicentre, phase III trial comparing the efficacy of CNP-00x (a virus-like particle that stimulates the immune system) and chemotherapy (arm B) with chemotherapy alone (arm A) for patients with unresectable, stage IIIb & IV non-small cell lung cancer (NSCLC). Eligible patients will be randomised 1:1 in the two arms. Treatment will last for 9 months.

Tumour assessments will be performed at baseline, and thereafter every 3 months (± 1 week) for the first year, and subsequently every 6 months (± 3 weeks) until death or first disease progression. In case of tumour progression detected in radiological imaging, patients will discontinue treatment and will be followed for their survival status.

Objective of the trial

The primary objective of the trial is to detect a 50% increase in the objective response rate (ORR) from 0.30 in the control to 0.45 in the experimental arm. Patients with a >30% decrease (as a best overall response) in the sum of diameters of all measurable lesion sites are considered to have achieved an objective response.

Best overall response (BOR) refers to the largest tumour size reduction from the baseline that was recorded throughout the trial. For example, if 55 out of the 100 patients of the one arm achieve an objective response at their *"best"* imaging, the ORR would be 55%.

In this superiority trial, two equally spaced looks are planned. A two-sided, asymmetric design was considered for the boundaries. The trial is powered by 80% with a one-sided type I error of 0.025.

Scenario A

Efficacy and futility bounds are derived using a Lan & DeMets's spending function approximating O'Brien & Fleming's bound.

Scenario B

Efficacy bound is derived using a Lan & DeMets's spending function approximating O'Brien & Fleming's bound. Futility bound is derived using the Hwang, Shih & De Cani's spending function with γ = -7.

ity of Piraeus University

Scenario C

Efficacy and futility bounds are derived using the Hwang, Shih & De Cani's spending function with γ_U = -3 and γ_L = -2, respectively.

Scenario	Lower bound	Upper bound
A	O'Brien&Fleming	O'Brien&Fleming
B	Hwang-Shih-DeCani (γ = -7)	O'Brien&Fleming
C	Hwang-Shih-DeCani (γ = -2)	Hwang-Shih-DeCani (γ = -3)

Table 5.4: Design parameters of the 3 scenarios.

Fixed design

The sample size of a trial having only one analysis would be 325 patients.

Results - Scenario A

In total, 343 patients have to be recruited to detect a 0.15 difference in ORR with 80% power. First analysis will be perfored after the results of 172 patients are available.

Stopping boundaries are shown in Figure 5.14, and they are also graphically depicted in Figure 5.15. In Figure 5.16, the black solid line represents the cumulative crossing probability at each interim vs. true treatment difference. Notice that above the value of -0.20, the probability of rejecting H_o rapidly decreases. Figure 5.17 presents the total expected sample size by true treatment difference.

```
> # BREATH Scenario A - Tabular Output
> gsBoundSummary(x, digits = 4)
  Analysis                    Value Efficacy Futility
 IA 1: 50%                        Z   2.9626   0.5594
    N: 172            p (1-sided)   0.0015   0.2879
                  delta at bound  -0.2183  -0.0412
              P(Cross) if delta=0   0.0015   0.7121
          P(Cross) if delta=-0.15   0.1770   0.0699
     Final                        Z   1.9686   1.9686
    N: 343            p (1-sided)   0.0245   0.0245
                  delta at bound  -0.1026  -0.1026
              P(Cross) if delta=0   0.0233   0.9767
          P(Cross) if delta=-0.15   0.8000   0.2000
```

Figure 5.14: Tabular summary of scenario A. An asymmetric, two-sided scheme was utilised for the boundaries. Efficacy and futility bounds were set using a Lan & DeMets's spending function approximating O'Brien & Fleming's bound.Treatment difference presented here is not a requirement, but an estimate of the difference required to cross each bound. P(Cross) is the probability of crossing the given bound (efficacy or futility) at or before the given analysis under the assumed treatment difference. Design assumes futility bound is discretionary (non-binding).

ity of Piraeus University

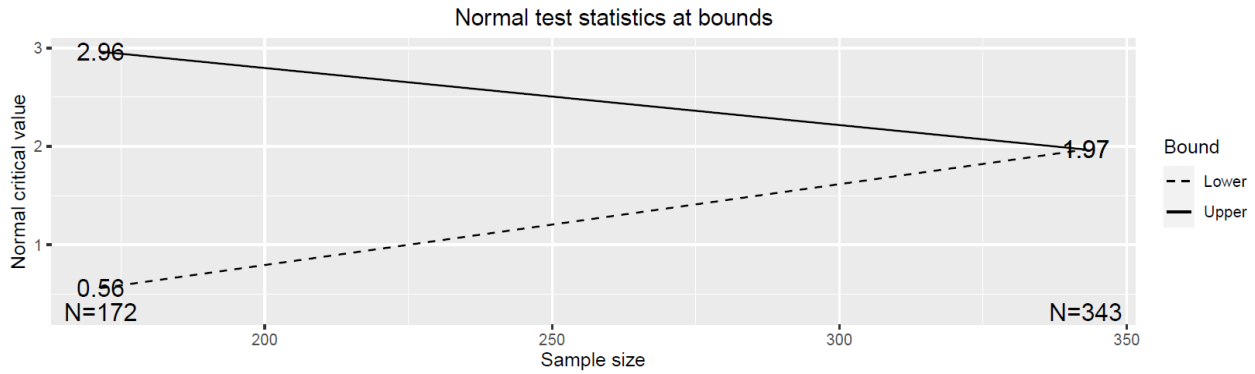


Figure 5.15: Stopping boundaries for scenario A. The solid (dashed) line represents the efficacy (futility) boundary.

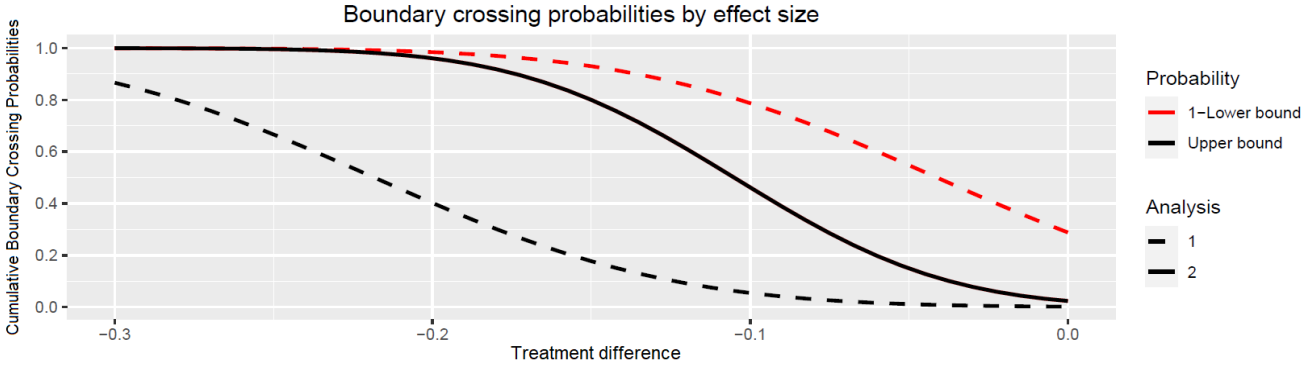


Figure 5.16: Cumulative boundary crossing probabilities by effect size for scenario A. Power by effect size is represented by the solid black line. Power at the first look is represented by the black dashed line. One minus the probability of crossing the futility bound by the first look is represented by the red dashed line.

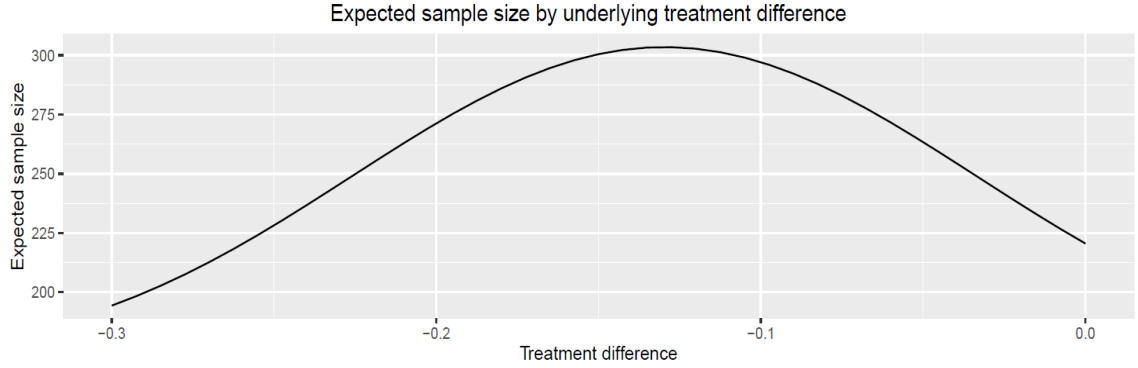


Figure 5.17: Total expected sample size by treatment difference for scenario A.

Results - Scenario B

In total, 327 patients have to be recruited to detect a 0.15 difference in ORR with 80% power. First analysis will be perfored after the results of 164 patients are available.

```
> # BREATH Scenario B - Tabular Output
> gsBoundSummary(x, digits = 4)
  Analysis                    Value Efficacy Futility
 IA 1: 50%                        Z   2.9626  -0.5348
    N: 164           p (1-sided)   0.0015   0.7036
                  delta at bound  -0.2238   0.0404
              P(Cross) if delta=0   0.0015   0.2964
         P(Cross) if delta=-0.15   0.1643   0.0059
     Final                        Z   1.9686   1.9686
    N: 327           p (1-sided)   0.0245   0.0245
                  delta at bound  -0.1052  -0.1052
              P(Cross) if delta=0   0.0250   0.9750
         P(Cross) if delta=-0.15   0.8000   0.2000
```

Figure 5.18: Tabular summary of scenario B. An asymmetric, two-sided scheme was utilised for the boundaries. Efficacy bound was set using a Lan & DeMets's spending function approximating O'Brien & Fleming's bound. Futility bound was set using the Hwang, Shih & De Cani's spending function with γ = -7. Treatment difference presented here is not a requirement, but an estimate of the difference required to cross each bound. P(Cross) is the probability of crossing the given bound (efficacy or futility) at or before the given analysis under the assumed treatment difference. Design assumes futility bound is discretionary (non-binding).

University of Piraeus

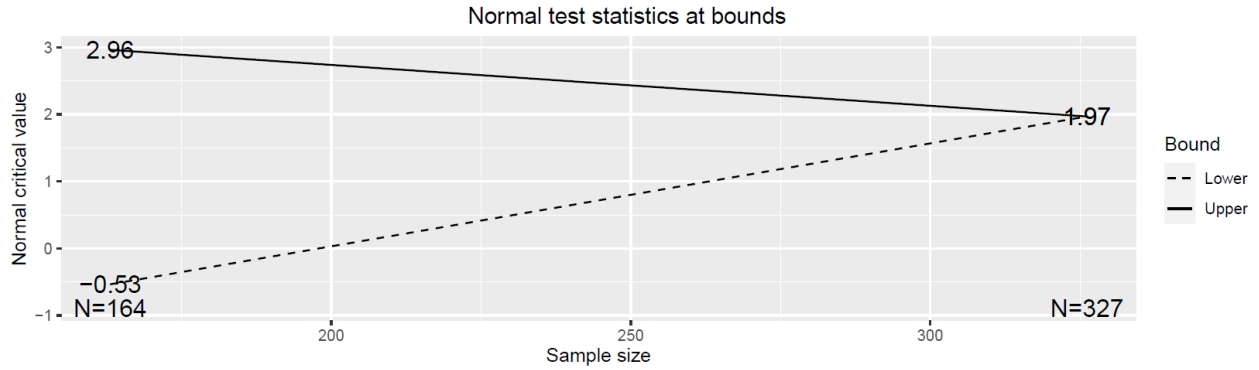


Figure 5.19: Stopping boundaries for scenario B. The solid (dashed) line represents the efficacy (futility) boundary.

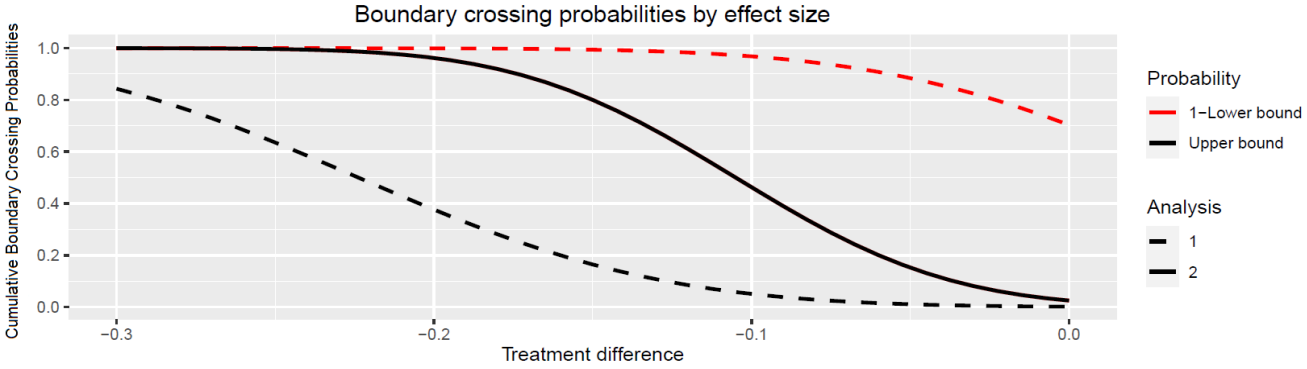


Figure 5.20: Cumulative boundary crossing probabilities by effect size for scenario B. Power by effect size is represented by the solid black line. Power at the first look is represented by the black dashed line. One minus the probability of crossing the futility bound by the first look is represented by the red dashed line.

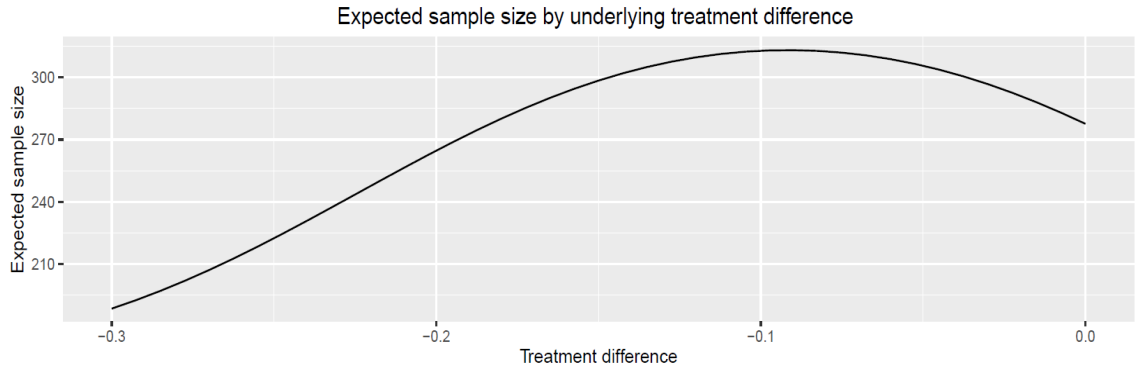


Figure 5.21: Total expected sample size by treatment difference for scenario B.

Results - Scenario C

In total, 341 patients have to be recruited to detect a 0.15 difference in ORR with 80% power. First analysis will be perfored after the results of 171 patients are available.

```
> # BREATH Scenario C - Tabular Output
> gsBoundSummary(x, digits = 4)
  Analysis                    Value Efficacy Futility
 IA 1: 50%                        Z   2.6075   0.4207
    N: 171           p (1-sided)   0.0046   0.3370
                  delta at bound  -0.1927  -0.0311
              P(Cross) if delta=0   0.0046   0.6630
         P(Cross) if delta=-0.15   0.2818   0.0538
     Final                        Z   1.9977   1.9977
    N: 341           p (1-sided)   0.0229   0.0229
                  delta at bound  -0.1044  -0.1044
              P(Cross) if delta=0   0.0240   0.9760
         P(Cross) if delta=-0.15   0.8000   0.2000
```

Figure 5.22: Tabular summary of scenario C. An asymmetric, two-sided scheme was utilised for the boundaries. Efficacy and futility bounds were set using the Hwang, Shih & De Cani's spending function with γ_U=-3 and γ_L=-2, respectively. Treatment difference presented here is not a requirement, but an estimate of the difference required to cross each bound. P(Cross) is the probability of crossing the given bound (efficacy or futility) at or before the given analysis under the assumed treatment difference. Design assumes futility bound is discretionary (non-binding).

University of Piraeus

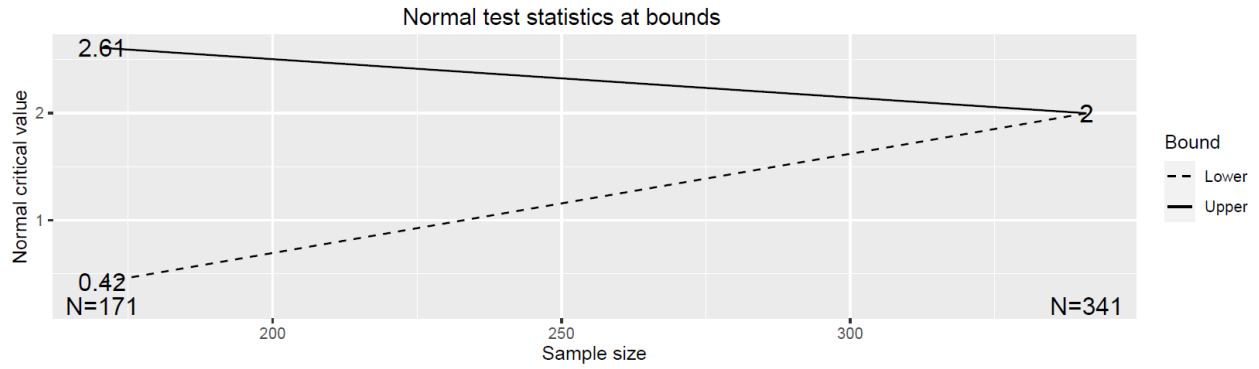


Figure 5.23: Stopping boundaries for scenario C. The solid (dashed) line represents the efficacy (futility) boundary.

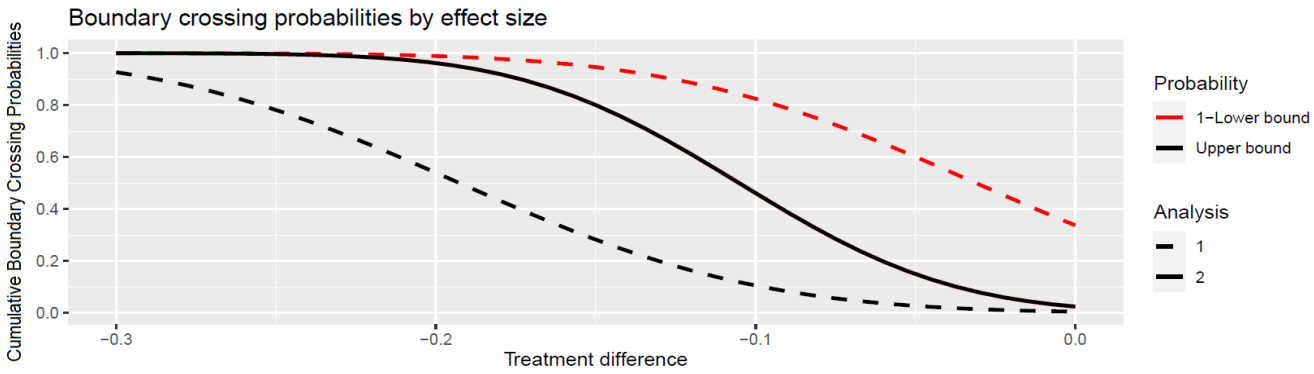


Figure 5.24: Cumulative boundary crossing probabilities by effect size for scenario C. Power by effect size is represented by the solid black line. Power at the first look is represented by the black dashed line. One minus the probability of crossing the futility bound by the first look is represented by the red dashed line.

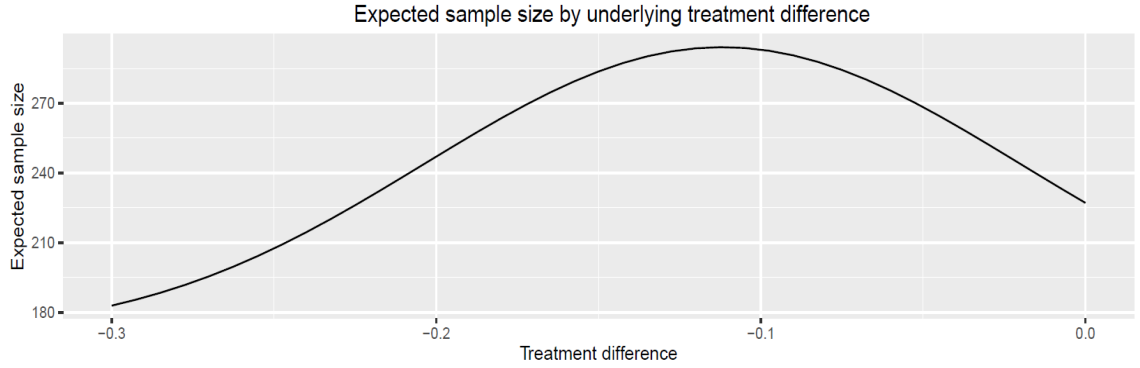


Figure 5.25: Total expected sample size by treatment difference for scenario C.

5.6 Discussion on BREATH trial

Scenarios A, B and C derive 343, 327 and 341 total sample size, respectively, while the sample size of the fixed design is 325 patients. Scenario A requires 18 more patients than the fixed design, which is translated in 5.5% increase; scenario B requires only 2 more patients (0.62% increase); scenario C requires 16 more patients (4.9% increase). Attained α is 0.023, 0.025 and 0.024 for A, B and C, respectively.

Scenarios B&C use the less conservative approach of Hwang, Shih & De Cani (see paragraph 2.3) to terminate the trial for futility. Specifically for design B which derives the smallest sample size of all 3 sequential designs, the lower futility boundary is the Hwang, Shih & De Cani's spending function with γ=-7. Such a small value of γ results in a quite slow spending rate of type II error (see Figure 5.27). Under the alternative, H_o will be rejected with 16% chance at the interim, engaging 164 patients of the total 327 required (a 49.5% decrease compared to the fixed design). Under H_o, the trial will be stopped for futility with 30% chance. Additionally, the model predicts that at the interim, the boundary will be crossed for efficacy (futility) if a difference of > 0.22 (< 0.040) is observed (Figure 5.18).

Scenarios A&C, require similar sample size and no more than 6% increase from the fixed design. For design A, both the upper and lower boundaries have been derived using an α-spending function approximating O'Brien & Fleming boundary. Under the alternative, H_o will be rejected with 18% chance at the interim, engaging 172 patients of the total 343 required (a 47.1% decrease compared to the fixed design). Under H_o, the trial will be stopped for futility with 71% chance. The respective probability for design B is only 30%, as cited above−the less conservative method that was used in scenario B resulted in a smaller sample size at the expense of having less change of stopping for futility under H_o.

Scenario C, on the other hand, uses the Hwang, Shih & De Cani's spending function for both boundaries, with the difference that in this case, γ is not so extreme. Under H_o, the trial will be stopped for futility with 66% chance. Under the alternative, H_o will be rejected with 28% chance at the interim−a value high enough to render scenario C the best choice to date.

Figures 5.26-5.28 show how β is spent for the 3 scenarios, while Figure 5.29 presents the Hwang, Shih & De Cani's spending function for various γ. Of note, with γ = -4 the function approximates the O'Brien & Fleming's bound, while with $\gamma = 1$ it approximates the Pocock's bound. The reader may now perceive the magnitude to which a value of γ = -7 diverges from conservatism.

ity of Piraeus University

Figure 5.26: Scenario A: Alpha & beta -spending functions (solid and dashed line, respectively) approximating O'Brien & Fleming's boundaries.

Figure 5.27: Scenario B: Alpha-spending function (solid line) approximating O'Brien & Fleming's bound. Beta-spending function (dashed line) is the Hwang, Shih & De Cani's spending function with γ = -7.

Figure 5.28: Scenario C: Alpha & beta-spending function (solid and dashed line, respectively) is the Hwang, Shih & De Cani's spending function with γ_U=-3 and γ_L=-2, respectively.

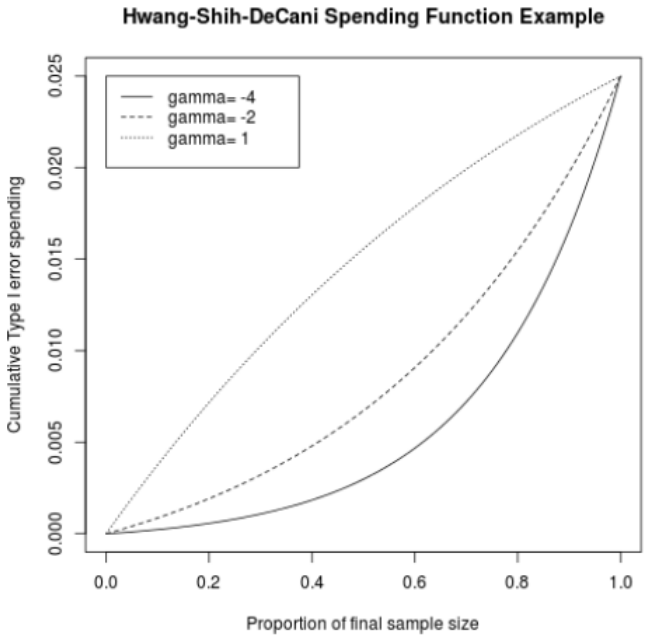


Figure 5.29: Hwang, Shih & De Cani's spending function for various γ. With γ =-4, the Hwang, Shih & De Cani's spending function approximates O'Brien & Fleming's bound; with γ =1, it approximates Pocock's bound. Source: https://rdrr.io/cran/gsDesign/

Figure 5.30 is a graphical representation of a design simulation. Crossing probabilities and sample size are given as a function of the Hwang, Shih & De Cani's γ, which is used for the lower bound. Upper bound is an a-spending function approximating O'Brien & Fleming. Notice that for γ =-7 we get design B.

If we demand a sample size of less than 340 patients, and >65% chance of stopping for futility under H_o, the scenario which derives the smallest total sample size is the one with γ = -2 (scenario D; result presented in Table 5.6). Although D engages 4 less patients than C, under H_1, the trial will be stopped for efficacy with much lower chance. Design C still remains the best choice.

Figure 5.31 presents crossing probabilities and sample size for various γ_L and γ_U. Both upper and lower boundaries were set using the Hwang, Shih & De Cani's spending function. The scenario which yields a sample size of less than 340 patients, >65% chance of stopping for futility under H_o, and >20% chance of stopping for efficacy under the alternative (and of them the smallest total sample size) is the one with (γ_L =-2, γ_U =-4.5) (scenario E). Although E requires 3 less patients than C (Table 5.6), it still has a lower chance of stopping for efficacy under H_1, rendering C is the best-case scenario.

University of Piraeus

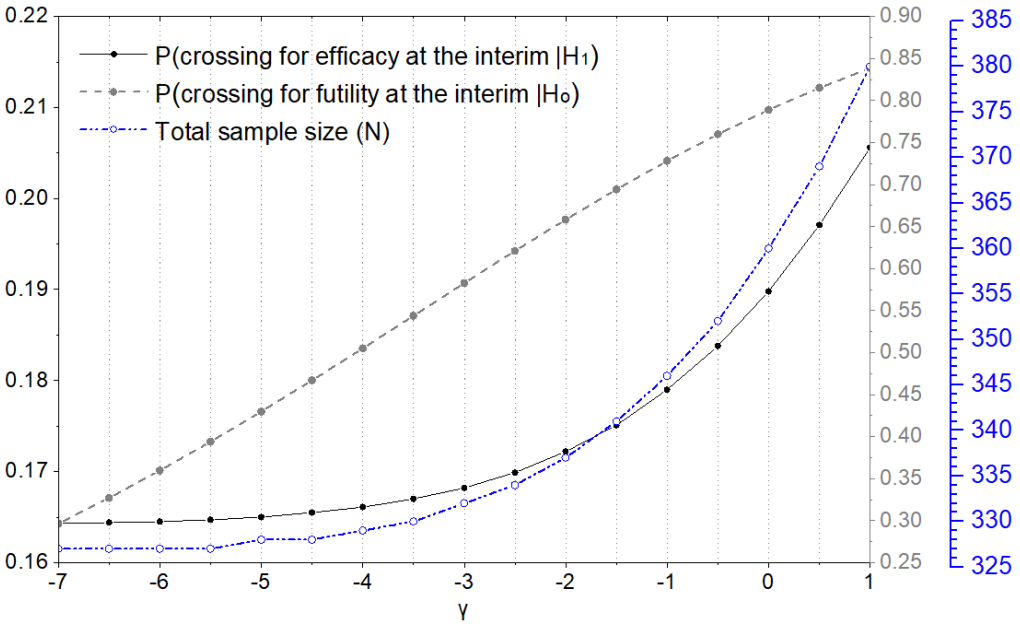


Figure 5.30: Crossing probabilities and sample size for various γ for a sequential design with two, equally-spaced looks. Upper boundary is an α-spending function approximating O'Brien & Fleming's bound; lower boundary is the Hwang, Shih & De Cani's spending function.

Scenario	Lower bound	Upper bound
A	O'Brien&Fleming	O'Brien&Fleming
B	Hwang-Shih-DeCani (γ = -7)	O'Brien&Fleming
C	Hwang-Shih-DeCani (γ = -2)	Hwang-Shih-DeCani (γ = -3)
D	Hwang-Shih-DeCani (γ = -2)	O'Brien&Fleming
E	Hwang-Shih-DeCani (γ = -2)	Hwang-Shih-DeCani (γ = -4.5)

Table 5.5: Design parameters of all 5 scenarios. Highlighted entries represent new scenarios.

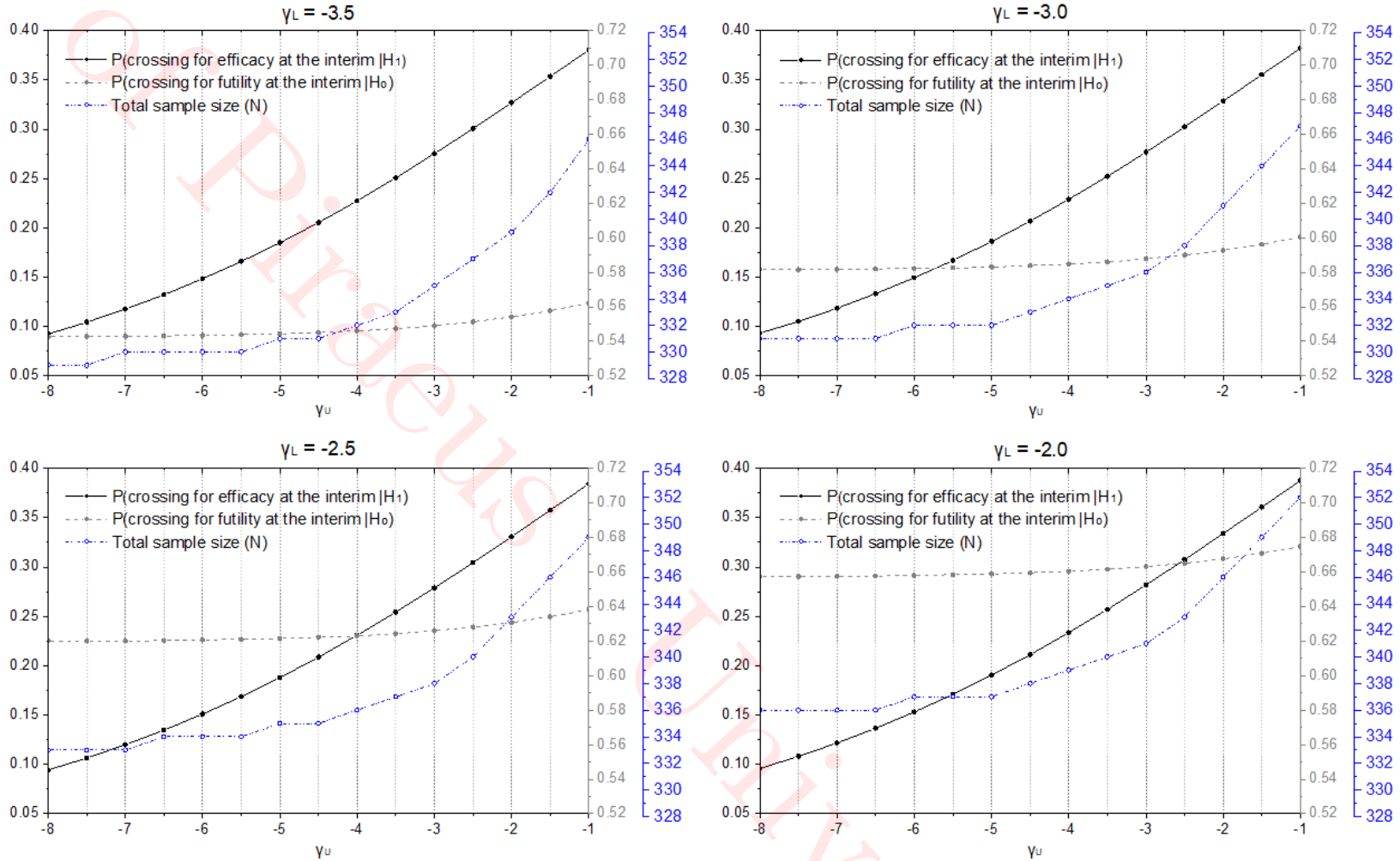


Figure 5.31: Crossing probabilities and sample size for various γ_L and γ_U for a sequential design with two, equally-spaced looks. Both upper and lower boundaries were set using the Hwang, Shih & De Cani's spending function.

City, University of London Institutional Repository: Pireaus

	Scenario					
	A	B	C	D	E	Fixed
N - Overall	343	327	341	337	338	325
n at the interim	172	164	171	169	169	*NA*
P(stop for efficacy at the interim$\mid H_o$)	0.0015	0.0015	0.0046	0.0015	0.0024	*NA*
P(stop for futility at the interim$\mid H_o$)	0.71	0.30	0.66	0.66	0.66	*NA*
P(stop for efficacy at the interim$\mid H_1$)	0.18	0.16	0.28	0.17	0.21	*NA*
P(stop for futility at the interim$\mid H_1$)	0.070	0.0059	0.054	0.054	0.054	*NA*

Table 5.6: Output design parameters and cumulative crossing probabilities of BREATH trial (all scenarios).

Conclusion on BREATH trial

The best-case scenario with regard to the total expected sample size and the crossing probabilities is scenario C. Scenario C utilises the Hwang, Shih & De Cani's spending function with γ_U=-3 and γ_L=-2 for the upper and the lower boundary, respectively.

5.7 Future research

A more sophisticated technique to determine the design parameters would be to use an optimization algorithm to approach the best solution in terms of sample size and crossing probabilities. For BREATH trial, an optimization method to determine γ_U and γ_L would probably be more appropriate. Also, designs utilising other spending functions (e.g. the exponential spending function) could yield even better results. Such cases are left for future research.

ity of Piraeus University

Review summary

In this review, we looked back into the late 1960s starting from recursive numerical integration by *Armitage et al.* [13]; *Pocock's (1977)* [43], *O'Brien & Fleming's (1979)* [37] and *Wang & Tsiatis's (1987)* [47] methods; error spending functions by *Lan & DeMets (1983)* [44] and *Hwang, Shih & De Cani (1990)* [54]; the *Slud & Wei's (1982)* [46] and *F-H-O (1984)* [49] methods; discrete sequential boundaries by *Bauer (1986)* [50]; stochastic curtailment [36]; various multi-stage designs by *Schaid et al. (1990)* [68], *Follman et al. (1994)* [67], *Stallard & Todd (2003)* [61], *Bischoff & Miller (2005)* [64]; seamless designs by *Stallard (2011)* [53]; the p-value combination test approach by *Bauer & Kisser (1999)* [72]; having more than one primary endpoint (*Bauer, 1991* [17]). Other methods, such as Bayesian posterior probabilities [83][84], the triangular and the double triangular test [85], symmetric [86] and asymmetric [87] designs, which are not presented here, have also shaped the landscape of statistical methodology for analysing accumulating data.

University of Piraeus

Bibliography

[1] R. Ng. *Drugs: From Discovery to Approval, Second Edition*. Copyright © 2009 John Wiley & Sons, Inc., 2008.

[2] S.W. Junod. FDA and clinical drug trials: A short history. *US Food & Drug Administration*.

[3] M. Clarke. The 1944 patulin trial of the British medical research council. *Journal of the Royal Society of Medicine*, 99, 2006.

[4] Unethical human experimentation. https://en.wikipedia.org/wiki/unethical_human_experimentation .

[5] US medical tests in Guatemala *"Crime Against Humanity"*. https://www.bbc.com/news/world-us-canada-11457552.

[6] P. Wandile and R. Ghooi. A role of ICH - GCP in clinical trial conduct. *Journal of Clinical Research and Bioethics*, 8(1), 2017.

[7] L. Alexander, A. Ivy, and H. Seibling. The Nuremberg Code. *Trials of war criminals before the Nuremberg military tribunals under Control Council Law*, 2(10):181 – 182, 1949.

[8] The Declaration of Helsinki. *Recommendations guiding doctors in clinical research adopted by the 18th World Medical Assembly, Helsinki, Finland*, 1964.

[9] ICH harmonised guideline integrated addendum to ICH E6(R1): Guideline for Good Clinical Practice ICH E6(R2). ICH Consensus Guideline. https://ichgcp.net/.

[10] Food and Drug Administration. The drug development process. https://www.fda.gov/patients/learn-about-drug-and-device-approvals/drug-development-process.

[11] M.A. McPherson. Statistics: The problem of examining accumulating data more than once. *Medical Intelligence*, 290:501 – 502, 1974.

University of Piraeus

[12] S.J. Pocock. Interim analysis for randomised clinical trials. The group sequential approach. *Biometrics*, 38:153 – 162, 1982.

[13] P. Armitage, C.K. McPherson, and B.C. Rowe. Repeated significance tests on accumulating data. *Journal of Royal Statistical Society*, Series A 132:235 – 244, 1969.

[14] C.K. McPherson and P. Armitage. Repeated significance tests on accumulating data when the null hypothesis is not true. *Journal of Royal Statistical Society*, 134(A):15 – 25, 1971.

[15] K.K.G. Lan and D.L. DeMets. Group sequential procedures: Calendar versus information time. *Statistics in Medicine*, 8:1191 – 1198, 1989.

[16] M.A. Proschan, D.A. Follmann, and M.A. Waclawiw. Effects on the assumption violations on type I error rate in group sequential monitoring. *Biometrics*, 48:1131 – 1143, 1992.

[17] P. Bauer. Multiple testing in clinical trials. *Statistics in Medicine*, 10:871 – 890, 1991.

[18] R.G. Miller. *Simultaneous statistical inference*. 2^{nd} edition, Springer, New York, 1981.

[19] E. Paulson. A sequential procedure for comparing several experimental categories with a standard or control. *Annals of Mathematical Statistics*, 33:438- – 443, 1962.

[20] E. Paulson. A sequential procedure for selecting the population with the largest mean from k normal populations. *Annals of Mathematical Statistics*, 35:174- – 180, 1964.

[21] C.W. Dunnett. A multiple comparisons procedure for comparing several treatments with a control. *Journal of American Statistical Association*, 60:573 – 583, 1965.

[22] R.E. Bechhofer, J. Kiefer, and M. Sobel. Sequential identification and ranking procedures. *University Chicago Press*, 1968.

[23] E. Spjøtvoll. On the optimality of some multiple comparison procedures. *Annals of Mathematical Statistics*, 48:398 – 411, 1972.

[24] S.A. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65 – 70, 1979.

[25] DeMets. D.L. and J.H. Ware. Group sequential methods in clinical trials with a one-sided hypothesis. *Biometrika*, 67:651 – 660, 1980.

[26] J.C. Hsu and D.G. Edwards. Sequential multiple comparison with the best. *Journal of American Statistical Association*, 78:958 – 964, 1983.

[27] R.J. Simes. Am improved bonferroni procedure for multiple tests of significance. *Biometrika*, 73:751 – 754, 1986.

[28] W.C. Kim, G. Stefansson, and J.C. Hsu. On confidence sets in multiple comparisons. *Statistical Decision Theory and Related Topics*, V, 1987.

[29] Y. Hochberg and A.C. Tamhane. *Multiple Comparison Procedures*. Wiley, New York, 1987.

[30] Y. Hochberg. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75:800 – 802, 1988.

[31] P. Bauer. On the assessment of the performance of multiple test procedures. *Biometrical Journal*, 29:895 – 906, 1987.

[32] P. Bauer. Sequential multiple testing for the elimination of inferior populations. *Proceedings of the Symposium in Gerolstein, Multiple Hypothesis Testing, Springer, Heidelberg*, 1988.

[33] P. Bauer. A sequential elimination procedure for choosing the best population(s) based on multiple testing. *Journal of Statistical Planning and Inference*, 21:245 – 252, 1989.

[34] P. Bauer. Sequential tests of hypotheses in consecutive trials. *Biometrical Journal*, 31:663 – 676, 1989.

[35] P. Bauer. Multistage testing with adaptive designs. *Biometric und Informatik in Medizin und Biologie*, 20:130 – 148, 1989.

[36] S.S. Emerson, J.M. Kittelson, and D.L. Gillen. On the use of stochastic curtailment in group sequential clinical trials. *UW Biostatistics Working Paper Series*, paper 243, 2005.

[37] P.C. O' Brien and T.R. Fleming. A multiple testing procedure for clinical trials. *Biometrics*, 35:549 – 556, 1979.

[38] A. Wald. *Sequential analysis*. John Wiley, New York, Chapman & Hall, London, 1947.

ity of Piraeus University

[39] H.F. Dodge and H.G. Roming. A method of sampling inspections. *Bell. System Technical Journal*, 8:613 – 631, 1929.

[40] C. Jennison and B. Turnbull. *Group sequential tests and repeated confidence intervals*. Handbook of sequential analysis (Ghosh Sen, eds), 1991.

[41] G. Wassmer. Basic concepts of group sequential and adaptive group sequential test procedures. *Statistical Papers*, 41:253 – 279, 2000.

[42] P. Bauer and K. Kohne. Evaluation of experiments with adaptive interim analyses. *Biometrics*, 50:1029 – 1041, 1994.

[43] S.J. Pocock. Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64:191 – 199, 1977.

[44] K.K.G. Lan and D.L. DeMets. Discrete sequential boundaries for clinical trials. *Biometrika*, 70:659 – 663, 1983.

[45] K. Kim and D.L. DeMets. Design and analysis of group sequential tests based on the type I error spending function. *Biometrika*, 74:149 – 154, 1987.

[46] E. Slud and L.J. Wei. Two sample repeated significance tests based on the modified Wilcoxon statistic. *Journal of American Statistical Association*, 77:862- – 868, 1982.

[47] S.K. Wang and A.A. Tsiatis. Approximately optimal one-parameter boundaries for group sequential trials. *Biometrika*, 43:193 – 199, 1987.

[48] A.A. Tsiatis. Repeated significance testing for a general class of statistics used in censored survival analysis. *Journal of American Statistical Association*, 77:855- – 861, 1982.

[49] T.R. Fleming, D.P. Harrington, and P.C. O' Brien. Designs for group sequential tests. *Controlled Clinical Trials*, 5:348 – 361, 1984.

[50] P. Bauer. Approximation of discrete sequential boundaries. *Biometrika*, 73:759- – 760, 1986.

[51] K.K.G. Lan, R. Simon, and M. Halperin. Stochastically curtailed tests in long-term clinical trials. *Communications in Statistics-Sequential Analysis*, 1:207- – 219, 1982.

[52] N. Stallard and K.M. Facey. Comparison of the spending function method and the Christmas tree correction for group sequential trials. *Journal of Biopharmaceutical Statistics*, 6:361 – 373, 1996.

[53] N. Stallard. Group-sequential methods for adaptive seamless phase II/III clinical trials. *Journal of Biopharmaceutical Statistics*, 21(4):787 – 801, 2011.

[54] I.K. Hwang, W.J. Shih, and J.S. DeCani. Group sequential designs using a family of type I error probability spending functions. *Statistics in Medicine*, 9:1439 – 1445, 1990.

[55] J.L. Haybittle. Repeated assessment of results in clinical trials of cancer treatment. *Br Journal of Radiology*, 44:793 – 797, 1971.

[56] C. Jennison and B.W. Turnbull. *Group sequential methods with applications to clinical trials*. Chapman & Hall, Boca Rato, London, New York, Washington, D.C., 1999.

[57] B.W. Turnbull. *Group sequential tests*. Encyclopedia of Statistical Sciences (Kotz, Read, Banks eds), 1997.

[58] J. Whitehead. *The design and analysis of sequential clinical trials*. Wiley: Chichester, 1997.

[59] J. Maca, S. Bhattacharya, V. Dragalin, P. Gallo, and M. Krams. Adaptive seamless phase ii/iii designs - background, operational aspects, and examples. *Drug Information Journal*, 40:463 – 473, 2006.

[60] P.F. Thall, R. Simon, and S.S. Ellenberg. A two-stage design for choosing among several experimental treatments and a control in clinical trials. *Biometrika*, 77:507 – 513, 1990.

[61] N. Stallard and S. Todd. Sequential designs for phase III clinical trials incorporating treatment selection. *Statistics in medicine*, 22:689 – 703, 2003.

[62] S. Todd and N. Stallard. A new clinical trial design combining phases 2 and 3: Sequential designs with treatment selection and a change of endpoint. *Drug Information Journal*, 39:109 – 118, 2005.

[63] M. Hellmich. Monitoring clinical trials with multiple arms. *Biometrics*, 57:892- – 898, 2001.

[64] W. Bischoff and F. Miller. Adaptive two-stage test procedures to find the best treatment in clinical trials. *Biometrika*, 92:197 – 212, 2005.

[65] R.E. Bechhofer. A single-sample multiple decision procedure for ranking means of normal populations with known variances. *Annals of Mathematical Statistics*, 25:16 – 39, 1954.

University of Piraeus

[66] D.G. Hoel and M. Mazumdar. An extension of identification and ranking procedures. paulson' s selection procedure. *Annals of Mathematical Statistics*, 39:2067 – 2074, 1968.

[67] D.A. Follman, M.A. Proschan, and N.L. Geller. Monitoring pairwise comparisons in multi-armed clinical trials. *Biometrics*, 50:325 – 336, 1994.

[68] D.J. Schaid, S. Wieand, and T.M. Therneau. Optimal two-stage screening designs for survival comparisons. *Biometrika*, 77:507 – 513, 1990.

[69] S.C. Chow and Y.H. Tu. On two-stage seamless adaptive design in clinical trials. *Journal of Formosan Medical Association*, 107(12*Suppl*):*S*52 – *S*60, 2008.

[70] S.C. Chow, M. Chang, and A. Pong. Statistical consideration of adaptive methods in clinical development. *Journal of Biopharmaceutical Statistics*, 15:575- – 591, 2005.

[71] R.A. Fisher. *Statistical methods for research workers*. Oliver & Boyd, Edinburgh, 4^{th} edition, 1932.

[72] P. Bauer and M. Kieser. Combining different phases in the development of medical treatments within a single trial. *Statistics in Medicine*, 18:1833 – 1848, 1999.

[73] N. Stallard and T. Fiede. Flexible group-sequential designs for clinical trials with treatment selection. *Statistics in Medicine*, 27:6209 – 6227, 2008.

[74] S.N. Roy. On a heuristic method of test construction and its use in multivariate analysis. *Annals of Mathematical Statistics*, 24:220 – 238, 1953.

[75] G. Hommel. Multiple test procedures for arbitrary dependence structures. *Metrika*, 33:321 – 336, 1986.

[76] B. Rüger. Das maximale signifikanzniveau des tests: Lehne H_o ab, wenn k unter n gegebenen tests zur ablehnung führen. *Metrika*, 25:171 – 178, 1978.

[77] G. Hommel. A stagewise rejective multiple test procedure based on a modified bonferroni test. *Biometrika*, 75:383 – 386, 1988.

[78] K. Abt. Descriptive data analysis: a concept between confirmatory and exploratory data analysis. *Methods of Information in Medicine*, 26:77 – 88, 1987.

[79] P. Bauer. Two stage sampling to simultaneously testing main and side effects in clinical trials. *Biometrics*, 28:871 – 879, 1986.

[80] S.D. Gupta. A note on some inequalities for the multivariate normal distribution. *Calcutta Statistical Association Bulletin*, 18:179 – 180, 1969.

[81] D. Sleplan. The one sided barrier problem for gaussian noise. *Bell Systems Technology*, 41:463 – 501, 1962.

[82] K. Kim and A.A. Tsiatis. Study duration for clinical trials with survival response and early stopping rule. *Biometrics*, 46(1):81 – 92, 1990.

[83] D.J. Spiegelhalter, L.S. Freedman, and M.K.B. Parmar. Bayesian approached to randomised trials. *Journal of Royal Statistical Society, Series A, General*, 157:357 – 387, 1994.

[84] D.F. Heitjan. Bayesian interim analysis of phase ii cancer clinical trials. *Statistics in Medicine*, 16:1791 – 1802, 1997.

[85] J. Whitehead and I. Stratton. Group sequential clinical trials with triangular continuation regions. *Biometrics*, 39:227 – 236, 1983.

[86] S.S. Emerson and T.R. Fleming. Symmetric group sequential test designs. *Biometrics*, 45:905 – 923, 1989.

[87] S. Pampallona and A.A. Tsiatis. Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favor of the null hypothesis. *Journal of Statistical Planning and Inference*, 42:19 – 35, 1994.

[88] D. Goldsman. Tutorial on indifference-zone normal means ranking and selection procedures. *Proceedings of the 1986 Winter Simulation Conference, J. Wilson, J. Henriksen, S. Roberts (eds.)*, 1986.

ity of Piraeus University

Appendix A

Ranking and Selection Procedures

The indifference zone approach−A section taken from a tutorial by Goldsman (1986) [88]

Let $\pi_1, \pi_2, ..., \pi_k$ be k normal populations with μ_i means ($i \in \{1, 2, ..., k\}$) and a common variance of σ^2.

Which of the k populations has the largest mean?

This question can be rephrased as:

Which population corresponds to the (k) *population with* $\mu_{(1)}, \mu_{(2)}, ..., \mu_{(k)}$ *denoting the ordered means?*

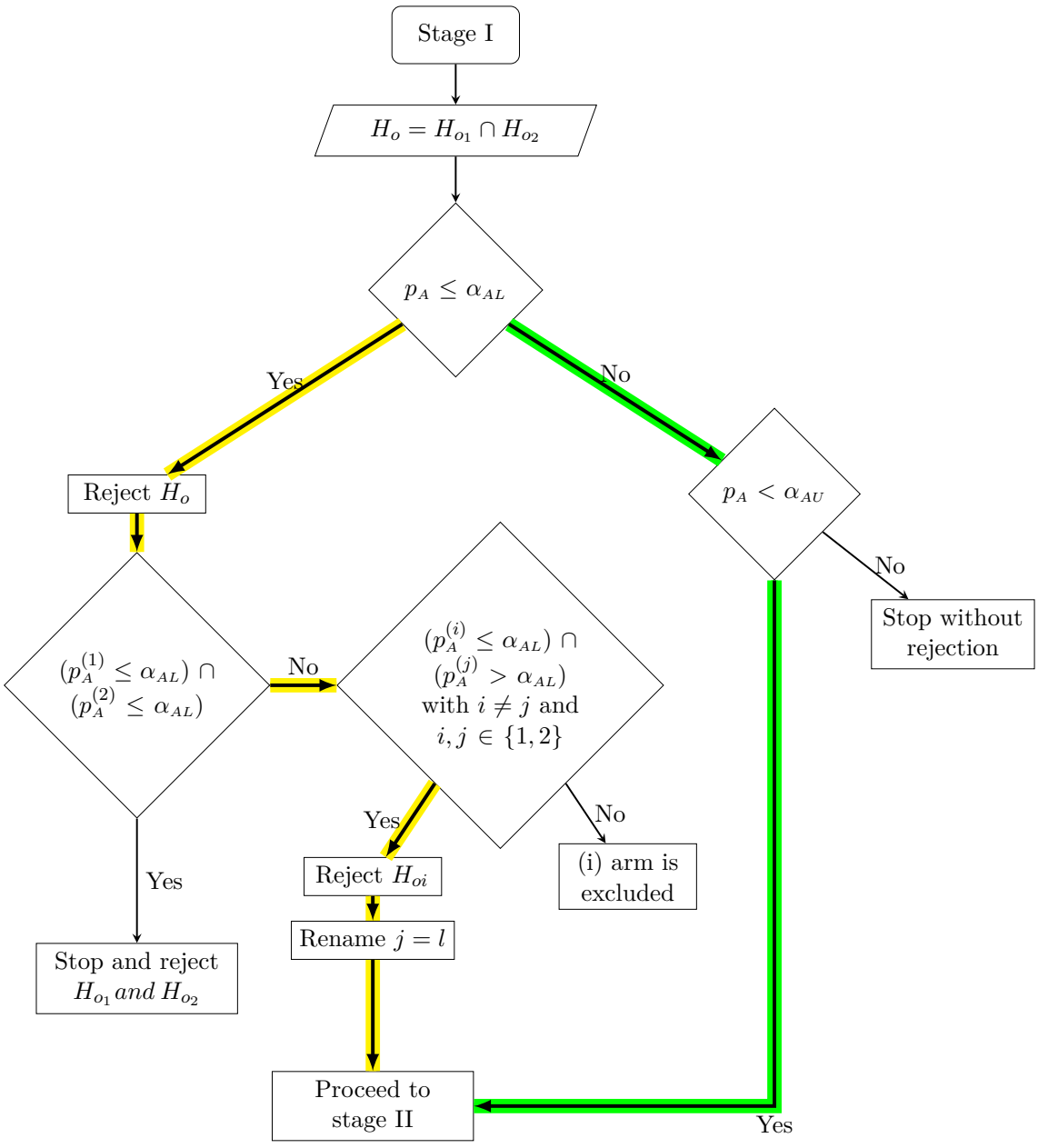
If the quantity of $\mu_{(k)} - \mu_{(k-1)}$ is very small, one could say that in practice, these two populations can be considered as one. A boundary δ is chosen as the smallest $\mu_{(k)} - \mu_{(k-1)}$ difference that the experimenter sees as *"worth detecting"*. So if $\mu_{(k)} - \mu_{(k-1)} \geq \delta$, the $\pi_{(k)}$ population will be chosen as the one with the largest mean. Whilst if $\mu_{(k)} - \mu_{(k-1)} < \delta$, the experimenter would be indifferent about which of the two populations to choose as the one with the largest mean.

$\Omega_\delta \equiv \{\mu \,|\, \mu_{(k)} - \mu_{(k-1)} \geq \delta\}$ is called the preference zone and its complement Ω_δ^* is called the indifference zone.

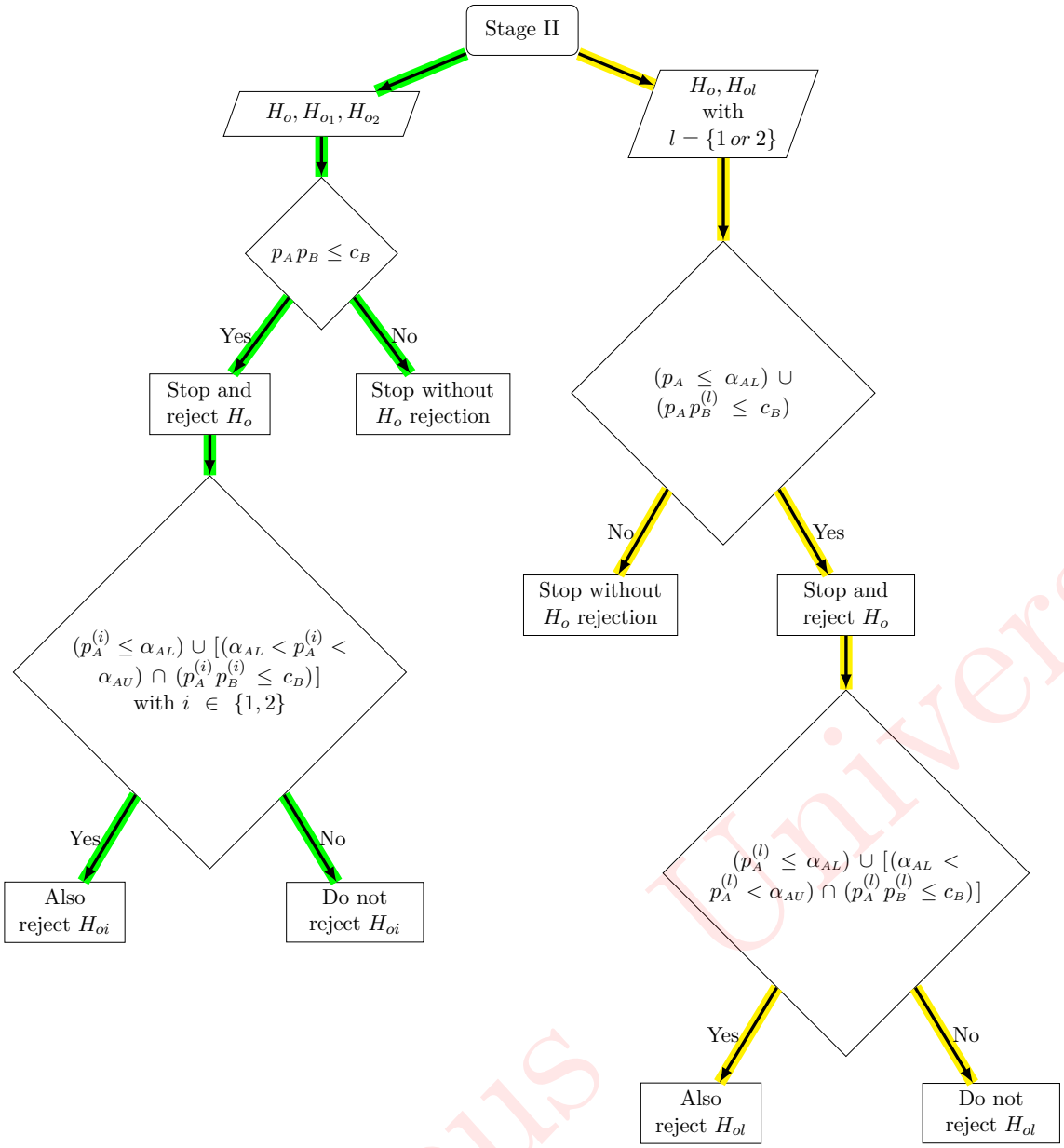
ity of Piraeus University

Appendix B

An adaptive two-stage design using the p-combination test approach



Multiple inference of the individual treatment-control comparisons in an adaptive seamless two-stage design having two investigational arms (and the control) and using the conservative approach of $p_A = 2min\,(p_A^{(1)}, p_A^{(2)})$ *- Stage I. [72]*



Stage II of the same design ($p_B = 2min(p_B^{(1)}, p_B^{(2)})$). *[72]*

University of Piraeus

Appendix C

R code for design simulations

```
# gsDesign package
install.packages("gsDesign")
library(gsDesign)

##########################################################
##################        ELPIDA TRIAL        ##################
##########################################################

# Fixed Design
nSurv(lambdaC = c(0.0459037867920494), hr = 0.65, hro = 1, eta = 0.05, gamma =
    c(5.5,11,16.5,22), R = c(2,2,2,6), S = NULL, T = NULL, minfup = 6, ratio = 1, alpha
    = 0.025, beta = 0.2)

# Group Sequential Design - Scenario A
x <- gsSurv(k = 2, test.type = 4, alpha = 0.025, beta = 0.2, astar = 0, timing = c(1),
    sfu = sfLDOF, sfupar = c(0), sfl = sfLDOF, sflpar = c(0), lambdaC =
    c(0.0459037867920494), hr = 0.65, hro = 1, eta = 0.05, gamma = c(5.5,11,16.5,22),
    R = c(2,2,2,6), S = NULL, T = NULL, minfup = 6, ratio = 1)

# Group Sequential Design - Scenario B
x <- gsSurv(k = 3, test.type = 4, alpha = 0.025, beta = 0.2, astar = 0, timing =
    c(0.30,0.60), sfu = sfLDOF, sfupar = c(0), sfl = sfLDOF, sflpar = c(0), lambdaC =
    c(0.0459037867920494), hr = 0.65, hro = 1, eta = 0.05, gamma = c(5.5,11,16.5,22),
    R = c(2,2,2,6), S = NULL, T = NULL, minfup = 6, ratio = 1)

# Group Sequential Design - Scenario C
x <- gsSurv(k = 3, test.type = 4, alpha = 0.025, beta = 0.2, astar = 0, timing =
    c(0.30,0.60), sfu = sfLDOF, sfupar = c(0), sfl = sfLDPocock, sflpar = c(0), lambdaC
    = c(0.0459037867920494), hr = 0.65, hro = 1, eta = 0.05, gamma =
    c(5.5,11,16.5,22), R = c(2,2,2,6), S = NULL, T = NULL, minfup = 6, ratio = 1)

# Group Sequential Design - Scenario D
x <- gsSurv(k = 2, test.type = 4, alpha = 0.025, beta = 0.2, astar = 0, timing = c(0.45),
    sfu = sfLDOF, sfupar = c(0), sfl = sfLDOF, sflpar = c(0), lambdaC =
    c(0.0459037867920494), hr = 0.65, hro = 1, eta = 0.05, gamma = c(5.5,11,16.5,22),
    R = c(2,2,2,6), S = NULL, T = NULL, minfup = 6, ratio = 1)

# Group Sequential Design - Scenario E
x <- gsSurv(k = 3, test.type = 4, alpha = 0.025, beta = 0.2, astar = 0, timing =
    c(0.35,0.60), sfu = sfLDOF, sfupar = c(0), sfl = sfLDOF, sflpar = c(0), lambdaC =
    c(0.0459037867920494), hr = 0.65, hro = 1, eta = 0.05, gamma = c(5.5,11,16.5,22),
```

University of Piraeus

```
    R = c(2,2,2,6), S = NULL, T = NULL, minfup = 6, ratio = 1)

# Tabular Output
gsBoundSummary(x, ratio = 1, digits = 4, tdigits = 1, timename = 'Month')

# Plot Design: Boundary
plot(x, plottype = 1, xlab = 'Events', ylab = 'Normal critical value')

# Plot Design: Power
plot(x, plottype = 2, xlab = 'Hazard ratio', ylab = 'Cumulative Boundary Crossing
    Probabilities')

# Plot Design: Conditional power at bounds
plot(x, plottype = 4, xlab = 'Events', ylab = expression(paste('Conditional power at ',
    theta, ' = ', hat(theta))))

# Plot Design: HR at bounds
plot(x, plottype = 8, xlab = 'Events', ylab = 'Estimated hazards ratio')

# Plot Design: Spending function
plot(x, plottype = 5, xlab = 'Proportion of final events', ylab = 'Proportion of
    spending')


##################################################################
################          BREATH TRIAL          ################
##################################################################

# Binomial Fixed Design
n <- nBinomial(p1 = 0.3, p2 = 0.45, deltao = 0, alpha = 0.025, beta = 0.2, ratio = 1)

# Group Sequential Design - Scenario A
x <- gsDesign(k = 2, test.type = 4, alpha = 0.025, beta = 0.2, astar = 0, timing = c(1),
    sfu = sfLDOF, sfupar = c(0), sfl = sfLDOF, sflpar = c(0), delta = 0, delta1 = -0.15,
    deltao = 0, endpoint = 'binomial', n.fix = n)

# Group Sequential Design - Scenario B
x <- gsDesign(k = 2, test.type = 4, alpha = 0.025, beta = 0.2, astar = 0, timing = c(1),
    sfu = sfLDOF, sfupar = c(0), sfl = sfHSD, sflpar = c(-7), delta = 0, delta1 = -0.15,
    deltao = 0, endpoint = 'binomial', n.fix = n)

# Group Sequential Design - Scenario C
x <- gsDesign(k = 2, test.type = 4, alpha = 0.025, beta = 0.2, astar = 0, timing = c(1),
    sfu = sfHSD, sfupar = c(-3), sfl = sfHSD, sflpar = c(-2), delta = 0, delta1 = -0.15,
    deltao = 0, endpoint = 'binomial', n.fix = n)

# Group Sequential Design - Scenario D
```

```
x <- gsDesign(k = 2, test.type = 4, alpha = 0.025, beta = 0.2, astar = 0, timing = c(1),
    sfu = sfLDOF, sfupar = c(0), sfl = sfHSD, sflpar = c(-2), delta = 0, delta1 = -0.15,
    deltao = 0, endpoint = 'binomial', n.fix = n)

# Group Sequential Design - Scenario E
x <- gsDesign(k = 2, test.type = 4, alpha = 0.025, beta = 0.2, astar = 0, timing = c(1),
    sfu = sfHSD, sfupar = c(-4.5), sfl = sfHSD, sflpar = c(-2), delta = 0, delta1 =
    -0.15, deltao = 0, endpoint = 'binomial', n.fix = n)

# Tabular Output
gsBoundSummary(x, digits = 4)

# Plot Design: Boundaries (Z)
plot (x, plottype = 1, xlab = 'Sample size', ylab = 'Normal critical value')

# Plot Design: Power
plot(x, plottype = 2, xlab = 'Treatment difference', ylab = 'Cumulative Boundary
    Crossing Probabilities')

# Plot Design: Conditional power at bounds
plot(x, plottype = 4, xlab = 'Sample size', ylab = expression(paste('Conditional power at
    ', theta, ' = ', hat(theta))))

# Plot Design: Expected samle size
plot(x, plottype = 6, xlab = 'Treatment difference', ylab = 'Expected sample size')

# Plot Design: Treatment Difference at Boundaries
plot(x, plottype = 3, xlab = 'Sample size', ylab = 'Estimated treatment difference at
    bounds')

# Plot Design: Spending function
plot(x, plottype = 5, xlab = 'Proportion of total sample size', ylab = 'Proportion of
    spending')



# R Version: 3.5.1
# gsDesign Version: 3.0.1
```

ity of Piraeus University